

# Stochastic thermodynamics based on incomplete information: Generalized Jarzynski equality with measurement errors with or without feedback

Christopher W. Wächtler<sup>1</sup>, Philipp Strasberg<sup>2</sup>, and Tobias Brandes

Institut für Theoretische Physik, Technische Universität Berlin, Hardenbergstr. 36,  
D-10623 Berlin, Germany

E-mail: <sup>1</sup> christopher.w.waechtler@campus.tu-berlin.de

E-mail: <sup>2</sup> phist@physik.tu-berlin.de

**Abstract.** In the derivation of fluctuation relations, and in stochastic thermodynamics in general, it is tacitly assumed that we can measure the system perfectly, i.e., without measurement errors. We here demonstrate for a driven system immersed in a single heat bath, for which the classic Jarzynski equality  $\langle e^{-\beta(W-\Delta F)} \rangle = 1$  holds, how to relax this assumption. Based on a general measurement model akin to Bayesian inference we derive a general expression for the fluctuation relation of the *measured* work and we study the case of an overdamped Brownian particle and of a two-level system in particular. We then generalize our results further and incorporate feedback in our description. We show and argue that, if measurement errors are fully taken into account by the agent who controls *and* observes the system, the standard Jarzynski-Sagawa-Ueda relation should be formulated differently. We again explicitly demonstrate this for an overdamped Brownian particle and a two-level system where the fluctuation relation of the measured work differs significantly from the efficacy parameter introduced by Sagawa and Ueda. Instead, the generalized fluctuation relation under feedback control,  $\langle e^{-\beta(W-\Delta F)-I} \rangle = 1$ , holds only for a superobserver having perfect access to *both* the system and detector degrees of freedom, independently of whether or not the detector yields a noisy measurement record and whether or not we perform feedback.

## 1. Introduction

During the last two decades we have seen an enormous progress in the understanding and description of the thermodynamic behaviour of small-scale systems, which are strongly fluctuating and arbitrary far from equilibrium. This includes, e.g., a consistent thermodynamic description at the single trajectory level and the discovery of so-called fluctuation relations which, in a certain sense, promote the status of the second law of thermodynamics from an inequality to an equality. A number of excellent review articles and monographs from different perspectives can be found in Refs. [1–7].

A tacit assumption underlying this framework, which is usually never discussed in any detail, is that we must be able to measure the stochastic trajectory  $\mathbf{z}(t)$  of a system perfectly, i.e., without measurement errors, in order to establish the framework of stochastic thermodynamics and to derive fluctuation relations. In practice, we know, however, that this is an experimental challenge for very small systems and, to put this thought even further, this might be the major obstacle in finding a fully satisfactory generalization of stochastic thermodynamics to quantum systems.

Extending the framework of stochastic thermodynamics to the case of incomplete or only partially available information has only recently attracted interest [8–13]. In our context, the results of García-García *et al.* [13], who have also derived a modified Jarzynski equality for faulty measurements, are of particular importance. Our results are indeed in agreement with their theory, though our point of view and derivation differs from them as we will discuss further in the main text below.

In addition, we also go one step beyond and include feedback based on faulty measurement results in our theory. In fact, the state of knowledge of the observer is of crucial importance in control theory and determines how “effective” the feedback control can be applied. However, if the experimentalist is forced to perform feedback based on faulty measurement results, it seems logical that she also uses the same (faulty) detector to infer other statistical properties of the system. Thus, we argue that, in order to extend stochastic thermodynamics to the case of feedback control *with* measurement errors, it is of crucial importance to take this measurement error consistently into account also during the time where no feedback is performed but where we still need to measure the system. This has indeed crucial consequences as we will examine below.

*Outline:* The article starts with a derivation of the standard Jarzynski equality (JE) based on a stochastic path integral method in order to establish the mathematical tools we will need in the following. Then, the rest of the article is divided into two main parts: Sec. 3 treats the case without feedback control and Sec. 4 the case with feedback control. In both cases we derive a general expression for the measured Jarzynski equality (MJE) of the measured work distribution for arbitrary measurement errors [Eqs. (20) and (35)]. In general, however, these might be extremely difficult to compute. Therefore, we present analytical results (underpinned by numerical simulations) for the two paradigmatic cases of an overdamped Brownian particle (OBP) in a harmonic potential and a two-level system (TLS). At all times we try to physically motivate our

results and shift most lengthy computations to the appendix. Furthermore, we comment on the use of mutual information in the JE in Sec. 5. Finally, in Sec. 6 we discuss our findings and point out to possible future applications.

## 2. Derivation of the Jarzynski equality for a driven system in a heat bath

Consider a system described by a Hamiltonian  $H_{\lambda(t)}(z)$ . Here,  $z$  might denote the position and momentum of a particle (i.e.,  $z = (x, p)$ ) or the discrete state of a system (such as spin up or down,  $z \in \{\uparrow, \downarrow\}$ ). The results derived below are independent of this consideration and we will use the notation of a continuous variable  $z$  most of the time. Next, suppose the system is in contact with a thermal bath at inverse temperature  $\beta$  and initially at  $t = 0$  in equilibrium with it, i.e.,  $p_{t=0}(z) = e^{-\beta H_{\lambda(0)}(z)} / Z_0$  with  $Z_0 = \int dz e^{-\beta H_{\lambda(0)}(z)}$ . Then, we change the Hamiltonian from  $t = 0$  to  $t = t_f$  as described by an arbitrary but fixed protocol  $\lambda(t)$ . Consequently, the work performed on the system,

$$W = W[\mathbf{z}] \equiv \int_0^{t_f} dt \dot{\lambda}(t) \frac{\partial H_{\lambda(t)}[\mathbf{z}(t)]}{\partial \lambda}, \quad (1)$$

along each trajectory  $\mathbf{z}(t) = \mathbf{z}$  becomes a stochastic quantity whose fluctuations are bounded by the following relation, which is also known as Jarzynski's equality (JE) [14, 15],

$$\langle e^{-\beta(W - \Delta F)} \rangle_{\mathbf{z}} = 1. \quad (2)$$

Here,  $\langle \dots \rangle_{\mathbf{z}}$  denotes an average over all possible system trajectories  $\mathbf{z}$  and  $\Delta F = -\beta^{-1}(\ln Z_f - \ln Z_0)$  denotes the change in equilibrium free energy. Eq. (2) can be derived in different ways and we will use stochastic path integrals and the method of time-reversed trajectories below.

In the formalism of stochastic path integrals the average of a trajectory-dependent quantity  $X[\mathbf{z}]$  can be expressed as [16]

$$\langle X[\mathbf{z}] \rangle_{\mathbf{z}} = \int \mathcal{D}[\mathbf{z}] \mathcal{P}[\mathbf{z}] X[\mathbf{z}] \quad (3)$$

where  $\mathcal{D}[\mathbf{z}]$  denotes a measure in the space of trajectories  $\mathbf{z}$  and  $\mathcal{P}[\mathbf{z}]$  the probability density (with respect to this measure) of choosing a trajectory  $\mathbf{z}$ . We now divide the time interval  $[0, t_f]$  into  $N$  time steps of duration  $\delta t = t_f / N$ . A particular trajectory  $\mathbf{z}$  is then approximated by its coordinates  $z_k \equiv z(t_k)$  at times  $t_k = k\delta t$ ,  $0 \leq k \leq N$ , such that

$$\mathbf{z}(t) \rightarrow [z_0, z_1, \dots, z_N] = \mathbf{z}. \quad (4)$$

Note that the limit  $N \rightarrow \infty$  by keeping  $t_f$  fixed is implied. The work along the trajectory is the discretized version of Eq. (1),

$$W[\mathbf{z}] = \sum_{k=1}^N (H_{\lambda_k}(z_{k-1}) - H_{\lambda_{k-1}}(z_{k-1})) \quad (5)$$

where  $\lambda_k$  denotes the value of the external control parameter at time  $t_k$ . Furthermore,

$$\int \mathcal{D}[\mathbf{z}] = \int dz_0 \int dz_1 \dots \int dz_N \quad (6)$$

where  $\int dz$  denotes an integral over a continuous variable (e.g., for an OBP) or a discrete sum (e.g., for a TLS). The probability density for a particular path is given by

$$\mathcal{P}[\mathbf{z}(t)] = p_{\lambda_0}(z_0) p_{\lambda_1}(z_0 \rightarrow z_1) p_{\lambda_2}(z_1 \rightarrow z_2) \dots p_{\lambda_N}(z_{N-1} \rightarrow z_N). \quad (7)$$

Here,  $p_{\lambda_0}(z_0)$  is the initial equilibrium distribution and  $p_{\lambda_k}(z_{k-1} \rightarrow z_k)$  denotes the transition probability from  $z_{k-1}$  to  $z_k$  in time  $\delta t$  where the driving protocol has the value  $\lambda_k$ . This factorization implicitly assumes Markovian system dynamics.

Of particular importance now will be the notion of a time-reversed path, denoted by  $\mathbf{z}^\dagger(t) = [z_N^*, z_{N-1}^*, \dots, z_0^*] = \mathbf{z}^\dagger$ , with time-reversed driving protocol  $\lambda^\dagger(t) = \lambda^*(t_f - t)$ .<sup>‡</sup> Here  $z_k^*$  indicates the time-reversal of  $z_k$ , e.g., if  $z_k = (x_k, p_k)$  for a particle with position  $x_k$  and momentum  $p_k$ , then  $z_k^* = (x_k, -p_k)$ . The probability density for such a path is

$$\mathcal{P}^\dagger[\mathbf{z}^\dagger] = p_{\lambda_N^*}(z_N^*) p_{\lambda_N^*}(z_N^* \rightarrow z_{N-1}^*) \dots p_{\lambda_1^*}(z_1^* \rightarrow z_0^*) \quad (8)$$

As usual in stochastic TD, we assume microreversibility (or local detailed balance) [17–19]

$$p_{\lambda_k^*}(z_k^* \rightarrow z_{k-1}^*) = p_{\lambda_k}(z_{k-1} \rightarrow z_k) e^{\beta \delta q_k(z_{k-1} \rightarrow z_k)} \quad (9)$$

where  $\delta q_k(z_{k-1} \rightarrow z_k) \equiv \delta q_k$  is the heat absorbed by the system during the time interval  $[t_{k-1}, t_k]$ . Due to normalization, we can write

$$\begin{aligned} 1 &= \int \mathcal{D}[\mathbf{z}^\dagger] \mathcal{P}^\dagger[\mathbf{z}^\dagger] \\ &= \int \mathcal{D}[\mathbf{z}^\dagger] \frac{p_{\lambda_N^*}(z_N^*)}{p_{\lambda_0}(z_0)} p_{\lambda_0}(z_0) p_{\lambda_1}(z_0 \rightarrow z_1) e^{\beta \delta q_1} \dots p_{\lambda_N}(z_{N-1} \rightarrow z_N) e^{\beta \delta q_N} \\ &= \int \mathcal{D}[\mathbf{z}] \mathcal{P}[\mathbf{z}] \frac{p_{\lambda_N^*}(z_N^*)}{p_{\lambda_0}(z_0)} e^{\beta(\delta q_1 + \dots + \delta q_N)} = \int \mathcal{D}[\mathbf{z}] \mathcal{P}[\mathbf{z}] \frac{p_{\lambda_N^*}(z_N^*)}{p_{\lambda_0}(z_0)} e^{\beta \delta q[\mathbf{z}]} \end{aligned} \quad (10)$$

where we used  $\mathcal{D}[\mathbf{z}^\dagger] = \mathcal{D}[\mathbf{z}]$  and introduced the heat  $\delta q[\mathbf{z}] \equiv \delta q_1 + \dots + \delta q_N$  absorbed along the full trajectory  $\mathbf{z}$ . Since the system is initially in equilibrium (in the forward as well as in the backward process), we have

$$\frac{p_{\lambda_N^*}(z_N^*)}{p_{\lambda_0}(z_0)} = \frac{Z_0}{Z_N} \exp[-\beta(H_{\lambda_N^*}(z_N^*) - H_{\lambda_0}(z_0))] \quad (11)$$

and furthermore  $H_{\lambda_N^*}(z_N^*) - H_{\lambda_0}(z_0) = H_{\lambda_N}(z_N) - H_{\lambda_0}(z_0) \equiv \Delta e(z_0, z_f)$ . By the first law of thermodynamics the energy difference between initial and final state along the

<sup>‡</sup> Note that in the presence of a magnetic field (or any other odd variable in the Hamiltonian) the sign of the field also changes under time-reversal.

trajectory is  $\Delta e(z_0, z_f) = q[\mathbf{z}] + W[\mathbf{z}]$ . Then, from Eq. (10) for  $N \rightarrow \infty$  (keeping  $t_f$  fixed) the original JE follows immediately:

$$1 = \int \mathcal{D}[\mathbf{z}] \mathcal{P}[\mathbf{z}] e^{\beta \Delta F} e^{-\beta W[\mathbf{z}]} = \langle e^{-\beta(W - \Delta F)} \rangle_{\mathbf{z}} . \quad (12)$$

To be precise and to emphasize that the statistical average  $\langle \dots \rangle_{\mathbf{z}}$  is taken over the system trajectories we explicitly use a subscript  $\mathbf{z}$ . This will change in the following.

### 3. Measured Jarzynski equality without feedback

Suppose now we measure the system coordinate  $z$  continuously with measurement outcome  $y$ , which in general can involve measurement errors and suppose the true system dynamics are inaccessible or hidden. Then the original JE, evaluated with the accessible measurement data, is in general not equal to unity, but depends on the difference of the true and measured work distribution.

More specifically, we introduce the conditional probability  $p_m(y|z)$  to obtain measurement outcome  $y$  given a particular state  $z$  of the system. The probability distribution of measurement outcomes  $y$  after a measurement is then

$$p'_m(y) = \int dz p_m(y|z) p(z) . \quad (13)$$

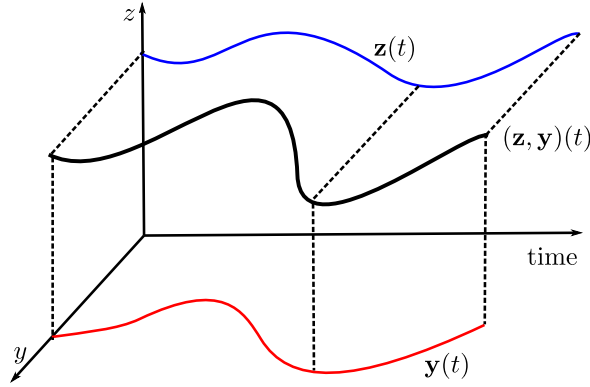
Given a particular measurement outcome  $y$ , the state of the system after the measurement is given by Bayes' rule and reads

$$p'(z|y) = \frac{p_m(y|z) p(z)}{p'_m(y)} . \quad (14)$$

The case of a perfect measurement, as usually considered in stochastic thermodynamics, is described by  $p_m(y|z) = \delta_{y,z}$  (where  $\delta_{y,z}$  denotes the Kronecker delta for a discrete state space or the Dirac distribution for a continuous system). It is then actually redundant to explicitly distinguish between the state of the system and the measurement result because  $p'_m(y) = p(z = y)$  and  $p'(z|y) = \delta_{y,z}$  (the final state is pure and coincides with the measurement result).

#### 3.1. General case

In order to incorporate the measurements on the system, we expand the phase space to the phase space of measured and true trajectories (see Fig. 1). A stochastic path in this extended space is denoted by  $(\mathbf{z}, \mathbf{y})$  and the probability of choosing such a path is simply denoted by  $\mathcal{P}[\mathbf{z}, \mathbf{y}]$ . The trajectory  $\mathbf{z}$  of the system is the projection of the whole trajectory onto the  $z$ -subspace and the probability distribution of this true stochastic path is given by  $\mathcal{P}[\mathbf{z}] = \int \mathcal{D}[\mathbf{y}] \mathcal{P}[\mathbf{z}, \mathbf{y}]$ . Equivalently, the measured trajectory  $\mathbf{y}$  lives in the  $y$ -subspace and its probability distribution is  $\mathcal{P}[\mathbf{y}] = \int \mathcal{D}[\mathbf{z}] \mathcal{P}[\mathbf{z}, \mathbf{y}]$ . Discretizing



**Figure 1.** Stochastic trajectory  $(\mathbf{z}, \mathbf{y})(t)$  (black) in the extended phase space of trajectories. The path  $\mathbf{z}(t)$  of the system (blue) is the projection onto the  $z$ -subspace and the measured trajectory  $\mathbf{y}(t)$  (red) is the projection onto the subspace of measurement. In general, measured and system trajectories are different.

the time interval  $[0, t_f]$  again into  $N$  time steps, the probability density of a path in the space of system and measured trajectories will be factorized as

$$\mathcal{P}[\mathbf{z}, \mathbf{y}] = p_{\lambda_0}(z_0, y_0) p_{\lambda_1}(z_0, y_0 \rightarrow z_1, y_1) \dots p_{\lambda_N}(z_{N-1}, y_{N-1} \rightarrow z_N, y_N) . \quad (15)$$

Our main assumptions are that the evolution of the system is independent of the measurement process and that the outcome of a measurement  $y_k$  only depends on the state of the system  $z_k$  at time  $t_k$ , i.e., we assume

$$p_{\lambda_k}(z_{k-1}, y_{k-1} \rightarrow z_k, y_k) = p_{\lambda_k}(z_{k-1} \rightarrow z_k) p_m(y_k | z_k) . \quad (16)$$

This can be seen as a Markov assumption for the measurement apparatus, i.e., the previous measurement result  $y_{k-1}$  does not influence the system evolution and the next measurement result. The conditional probability  $p_m(y_k | z_k)$  quantifies the uncertainty of the measurement (see Eqs. (13) and (14)).

The measured work  $W_m[\mathbf{y}]$  along a measurement trajectory  $\mathbf{y}$  is defined as in Eqs. (1) and (5) by interchanging  $\mathbf{z}$  with  $\mathbf{y}$  and is in general *different* from the true work  $W = W[\mathbf{z}]$ . Even on average it might be that  $\langle W_m[\mathbf{y}] \rangle_{\mathbf{y}} \neq \langle W[\mathbf{z}] \rangle_{\mathbf{z}}$ . Nevertheless, we assume that the Hamiltonian of the system is known to us and unchanged by the measurement; the only mistake is in the measurement outcome  $y$  (see Ref. [13] for the case of different Hamiltonians).

From an experimental point of view it only makes sense to consider the distribution of measured work and we may write the average of the exponential of measured work and free energy difference  $\Delta F$  as

$$\begin{aligned} \langle e^{-\beta(W_m - \Delta F)} \rangle_{\mathbf{y}} &= \int \mathcal{D}[\mathbf{y}] \mathcal{P}[\mathbf{y}] e^{-\beta(W_m[\mathbf{y}] - \Delta F)} = \int \mathcal{D}[\mathbf{y}] \mathcal{D}[\mathbf{z}] \mathcal{P}[\mathbf{z}, \mathbf{y}] e^{-\beta(W_m[\mathbf{y}] - \Delta F)} \\ &= \int \mathcal{D}[\mathbf{y}] \mathcal{D}[\mathbf{z}] \mathcal{P}[\mathbf{z}] \prod_{i=0}^N p_m(y_i | z_i) e^{-\beta(W_m[\mathbf{y}] - \Delta F)} \end{aligned} \quad (17)$$

where in the last step we used (16). Again the assumption of microreversibility (see Eq. (9)) allows us to write Eq. (17) as

$$\langle e^{-\beta(W_m - \Delta F)} \rangle_{\mathbf{y}} = \int \mathcal{D}[\mathbf{y}] \mathcal{D}[\mathbf{z}] \mathcal{P}^\dagger[\mathbf{z}^\dagger] \prod_{i=0}^N p_m(y_i | z_i) e^{-\beta \Delta e^\dagger(z_0, z_f)} e^{\beta \delta q^\dagger[\mathbf{z}^\dagger]} e^{-\beta W_m[\mathbf{y}]} . \quad (18)$$

Here,  $\Delta e^\dagger(z_0, z_f)$  and  $\delta q^\dagger[\mathbf{z}^\dagger]$  are the energy difference and the exchange of heat with the reservoir along the system's backward trajectory, respectively. The first law also holds for the backwards paths of the system,  $\Delta e^\dagger = -\Delta e(z_0, z_f) = W^\dagger[\mathbf{z}^\dagger] + \delta q^\dagger[\mathbf{z}^\dagger]$  and assuming time-reversal symmetry of the measurement,  $p_m(y_i | z_i) = p_m(y_i^* | z_i^*)$ , we can further simplify Eq. (18) to

$$\begin{aligned} \langle e^{-\beta(W_m - \Delta F)} \rangle_{\mathbf{y}} &= \int \mathcal{D}[\mathbf{y}] \mathcal{D}[\mathbf{z}] \mathcal{P}^\dagger[\mathbf{z}^\dagger] \prod_{i=0}^N p_m(y_i^* | z_i^*) e^{-\beta W^\dagger[\mathbf{z}^\dagger]} e^{\beta W_m^\dagger[\mathbf{y}^\dagger]} \\ &= \int \mathcal{D}[\mathbf{y}] \mathcal{D}[\mathbf{z}] \mathcal{P}^\dagger[\mathbf{z}^\dagger, \mathbf{y}^\dagger] e^{\beta(W_m^\dagger[\mathbf{y}^\dagger] - W^\dagger[\mathbf{z}^\dagger])} \end{aligned} \quad (19)$$

where we have used that the measured work is asymmetric under time reversal,  $W_m^\dagger[\mathbf{y}^\dagger] = -W_m[\mathbf{y}]$ , which directly follows from the corresponding property of the true work. Thus, one finally arrives at the following expression for the MJE:

$$\langle e^{-\beta(W_m - \Delta F)} \rangle_{\mathbf{y}} = \left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} . \quad (20)$$

This expression results from a formal manipulation and is at this point, however, still explicitly dependent on the (backward) trajectories  $\mathbf{z}^\dagger$  of the system and is therefore of limited practical use. Later on we will see how to overcome this difficulty for various examples where we use Eq. (20) as our formal starting point. Note that depending on the probability distribution  $p(W^\dagger, W_m^\dagger)$  an expansion in terms of the moments of the distribution could be also attempted.

As an important limiting case we immediately see that for a perfect measurement,  $p_m(y_k | z_k) = \delta_{y_k, z_k}$ , the measured work coincides with the work of the system,  $W_m[\mathbf{y}] = W[\mathbf{z}]$ , and the right hand side becomes unity recovering the original JE (see Eq. (2)). Moreover, the right hand side of Eq. (20) may also be equal to one if there is a certain symmetry in the driven system, such that  $W_m^\dagger[\mathbf{y}^\dagger] = W^\dagger[\mathbf{z}^\dagger]$  (see, e.g., Sec. 3.2).

Finally, let us comment on recent work by García-García *et al.* [13], who also derive a modified JE including measurement errors and which is equivalent to our result, Eq. (20). However, their point of view as well as the derivation differ from the present approach. García-García *et al.* introduce the error  $E[\mathbf{z}, \mathbf{y}] = W[\mathbf{z}] - W_m[\mathbf{y}]$  of system and measured work and derive a fluctuation theorem for the joint distribution of the measured work and this error [13]:

$$\ln \frac{p'(W_m, E)}{p^\dagger(-W_m, -E)} = \beta(W_m + E - \Delta F) . \quad (21)$$

From the latter relation, one can immediately derive Eq. (20). Thus, whereas all measurement errors in Ref. [13] are incorporated at the level of the final work distribution



$p'(W_m, E)$ , we **start with a particular measurement model for the state of the system expressed in terms of  $p_m(y_k|z_k)$** . This is closer to a microscopic modeling of the situation because any measurement model for the system  $p_m(y_k|z_k)$  will also yield a certain work distribution  $p'(W_m, E)$ , whereas for a given work distribution  $p'(W_m, E)$  there might be many different measurement models (and even different systems) which yield the same  $p'(W_m, E)$ . Thus, our findings show a completely different path to derive fluctuation theorems in the presence of measurement errors. Whether our approach or the one of Ref. [13] is superior might depend strongly on the specific situation and the system under study.

In the following sections we examine two paradigmatic systems for which the right hand side of Eq. (20) can be evaluated analytically, namely an overdamped Brownian particle (OBP) in a harmonic potential in Sec. 3.2 and a two-level system (TLS) in Sec. 3.3.

### 3.2. Overdamped Brownian motion

We consider the overdamped dynamics of a particle in a harmonic potential in one dimension such that the Hamiltonian of the system is only given by the potential energy:

$$H_{\lambda(t)}(z) = V_{\lambda(t)}(z) = f_{\lambda(t)}(z - \mu_{\lambda(t)})^2. \quad (22)$$

The stiffness  $f_{\lambda(t)}$  as well as the center of the potential  $\mu_{\lambda(t)}$  can be altered in time by an external driving protocol  $\lambda(t)$ . To simulate the system dynamics we use the Langevin equation

$$\dot{z}(t) = -\beta D V'_{\lambda(t)}(z) + \sqrt{2D} \xi(t) \quad (23)$$

with diffusion constant  $D$ , which is related to the friction constant  $\gamma$  by the Einstein relation  $D = (\beta\gamma)^{-1}$ , and Gaussian white noise  $\xi(t)$ .

We specify our measurement model by assuming that the measured position of the particle  $y_i$  is normally distributed around the real position  $z_i$  with a standard deviation of  $\sigma_m$ ,

$$p_m(y_i|z_i) = \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left(-\frac{(z_i - y_i)^2}{2\sigma_m^2}\right), \quad (24)$$

such that, if  $\sigma_m \rightarrow 0$ , the conditional probability becomes a Dirac distribution and the measured coordinate coincides with the true coordinate of the particle. Such a Gaussian measurement model might be a good approximation for a noisy measurement without systematic error (i.e., we have  $\langle y \rangle = \langle z \rangle$ ) and simplifies a lot analytical calculations.

Note that the Langevin equation (23) now merely presents a convenient numerical tool. From the point of view of the observer, it has no objective reality unless  $\sigma_m = 0$ . The correct state of knowledge of the observer would be indeed described by a stochastic Fokker-Planck equation [20, 21].



*Continuous driving protocol.* Evaluating the general expression, Eq. (20), for a continuous and piecewise differentiable (c.p.d.) driving protocol  $\lambda(t)$  yields (see Appendix A.1 for the derivation)

$$\langle e^{-\beta(W_m - \Delta F)} \rangle_y = e^{-\sigma_m^2 \beta \Delta f} \quad (25)$$

where  $\Delta f \equiv f_{\lambda(t_f)} - f_{\lambda(0)}$ . The right hand side of the above equation equals unity for  $\sigma_m = 0$  corresponding to the original JE. Similarly, if we vary the width of the potential periodically such that  $f_{\lambda(0)} = f_{\lambda(t_f)}$ , then the original JE is also recovered. However, this attribute is, as far as we know, specific to the model of the overdamped Brownian particle with c.p.d. driving protocol. In general the right hand side will be different from one. Interestingly, shifting the center  $\mu_{\lambda(t)}$  of the potential has no effect at all on the MJE. Furthermore, if we define an effective free energy,  $\Delta \tilde{F} \equiv \Delta F + \sigma_m^2 \Delta f$ , which may be interpreted as an additional contribution due to the uncertainty of the measurements, a JE of the form  $\langle e^{-\beta(W_m - \Delta \tilde{F})} \rangle_y = 1$  holds.

*Instantaneous change of driving protocol ("quench").* We also derive in Appendix A.2 an analytic expression for the MJE for an instantaneous change of the system Hamiltonian at a time  $t_m$  (also called a 'quench'). We consider here that the position and the width of the parabola is altered at the same time and is constant before and after  $t_m$ . We find

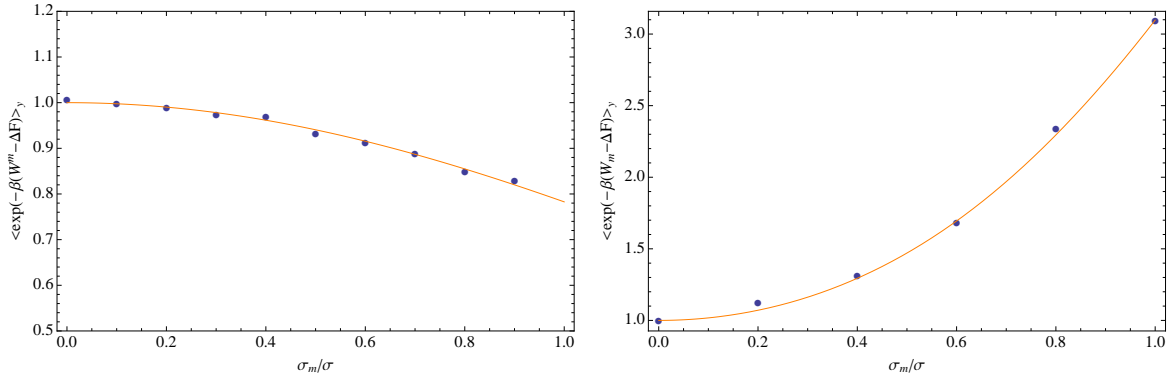
$$\langle e^{-\beta(W_m - \Delta F)} \rangle_y = \frac{1}{\sqrt{1 + 2\beta\sigma_m^2 \frac{f_{\lambda(0)}}{f_{\lambda(t_f)}} \Delta f}} \exp \left\{ \frac{2\beta^2 \sigma_m^2 f_{\lambda(0)}^2 \Delta \mu^2}{1 + 2\beta^2 \sigma_m^2 \frac{f_{\lambda(0)}}{f_{\lambda(t_f)}} \Delta f} \right\} \quad (26)$$

where  $\Delta \mu \equiv \mu_{\lambda(t_f)} - \mu_{\lambda(0)}$  is the difference of the center of the parabola before and after  $t_m$  §

*Numerics.* In order to verify our findings we performed Brownian dynamics (BD) simulations and used the *weighted ensemble path sampling* algorithm [22], which shifts the computational resources towards the sampling of rare trajectories, which have the largest impact on the JE. It has been shown that this method is statistically exact for a broad class of Markovian stochastic processes [23]. Please note that we set  $\beta \equiv 1$  for all simulations in this paper.

As a simple example we change both parameters of the potential continuously and linearly in time. We choose  $f_{\lambda(t)} = f_{\lambda(0)} + \alpha t$  and  $\mu_{\lambda(t)} = \mu_{\lambda(0)} + \alpha' t$ . For this driving scheme we find very good agreement of BD simulation and the analytic expression, Eq. (25), which is presented in Fig. 2 (left).

§ As a side remark note that Eq. (25) cannot be reproduced from Eq. (26) although a quench can be modeled as a limit of a series of continuous functions. This has nothing to do with the phenomenon of absolute irreversibility [?]. Instead, from our derivation in Appendix A.1 it becomes apparent that this procedure would require us to interchange the limit of the series of continuous functions with an integral, which is only allowed for a *uniformly convergent* series, but a series of continuous functions converging to a quench (which is not continuous) is not uniformly convergent (but pointwise instead).



**Figure 2.** Left: Comparison of BD simulation (marks) and the analytic expression (Eq. (25), line) where the system is driven continuously and we choose  $f_{\lambda(0)} = 4.0$ ,  $\alpha = 2.0$ ,  $\mu_{\lambda(0)} = 0.0$  and  $\alpha' = 2.0$ . Right: Results of simulation and the analytic expression for a quench of the OBP (see Eq. (26)), where  $f_{\lambda(0)} = 2.0$ ,  $f_{\lambda(t_f)} = 4.0$ ,  $\mu_{\lambda(0)} = 0.0$  and  $\mu_{\lambda(t_f)} = 1.0$ . The quench is performed at time  $t_m = t_f/2$ . For both driving schemes  $t_f = 5.0$ ,  $D = 2.0$  and we choose  $\Delta t = 0.0001$ .

Furthermore, we compare Eq. (26) with simulation results where initially the Hamiltonian of the system is given by  $H_0 = f_{\lambda(0)}(z - \mu_{\lambda(0)})^2$  and which is instantaneously changed to  $H_f = f_{\lambda(t_f)}(z - \mu_{\lambda(t_f)})^2$  at  $t_m$ . In Fig. 2 (right) we show the results of the BD simulation (marks) as well as the analytic expression (line) for different values of  $\sigma_m$  verifying our findings also for a quench.

### 3.3. Two-level system

Consider a driven system consisting of two energy levels, a ground state with energy  $\varepsilon_{\lambda(t)}(g)$  and an excited state with energy  $\varepsilon_{\lambda(t)}(e)$ , coupled to a heat bath with inverse temperature  $\beta$ . The master equation (ME) describing this system is

$$\frac{d}{dt} \begin{pmatrix} p_g(t) \\ p_e(t) \end{pmatrix} = \Gamma \begin{pmatrix} -e^{-\beta\omega_{\lambda(t)}/2} & e^{\beta\omega_{\lambda(t)}/2} \\ e^{-\beta\omega_{\lambda(t)}/2} & -e^{\beta\omega_{\lambda(t)}/2} \end{pmatrix} \begin{pmatrix} p_g(t) \\ p_e(t) \end{pmatrix} \quad (27)$$

Here, we denoted the energy gap of excited and ground state by  $\omega_{\lambda(t)} \equiv \varepsilon_{\lambda(t)}(e) - \varepsilon_{\lambda(t)}(g)$  and  $p_{g/e}(t)$  denotes the probability to find the system in the ground/excited state.

We measure the state of the system continuously with  $(1-\eta)$  being the probability of measuring the state of the system correctly and consequently  $\eta$  of measuring it wrongly, i.e., we set  $p_m(y_k|z_k) = (1-\eta)\delta_{y_k,z_k} + \eta(1-\delta_{y_k,z_k})$  with  $\eta \in [0, 1]$ .

*Continuous driving protocol.* The MJE of the TLS, where the external control parameter  $\lambda(t)$  is c.p.d., can be well approximated by (see Appendix A.3 for the derivation)

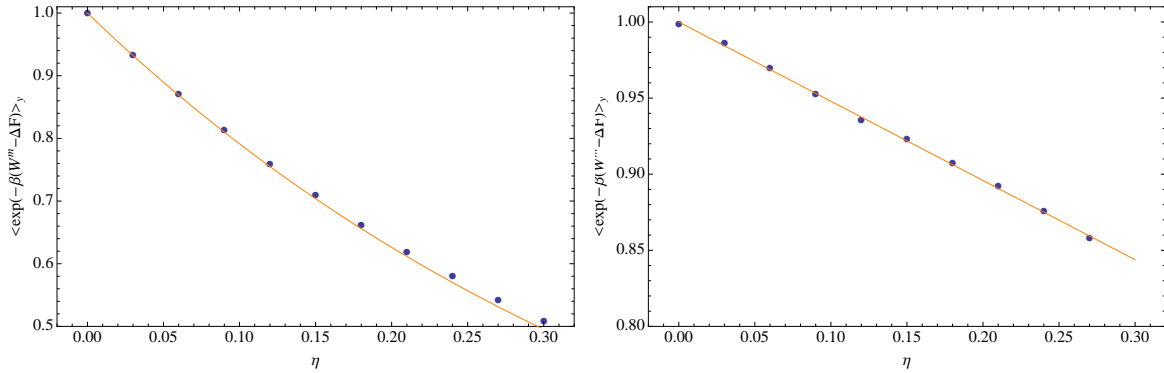
$$\langle e^{-\beta(W^m - \Delta F)} \rangle_{\mathbf{y}} \approx \exp \left( -\eta\beta \int_0^{t_f} dt \dot{\omega}_{\lambda^\dagger(t)} (p_e^\dagger(t) - p_g^\dagger(t)) \right) \quad (28)$$

where  $p_{g/e}^\dagger(t)$  denotes the probability that the system is in the ground/excited state in the backward process at time  $t$ , respectively. Furthermore,  $\omega_{\lambda^\dagger(t)}$  denotes the energy gap of the TLS. We remark, that for a c.p.d. protocol with nondifferentiable points at  $0 < t_1 < \dots < t_K < t_f$  we have to split the integral at the respective points as  $\int_0^{t_f} dt = \int_0^{t_1} dt + \int_{t_1}^{t_2} dt + \dots + \int_{t_K}^{t_f} dt$ . Moreover, Eq. (28) is exact up to first order in  $\eta$ . For higher orders (say  $\eta^k$ ) we have to assume that  $\mathcal{P}[z_{i_1}, \dots, z_{i_k}] \approx p(z_{i_1}) \dots p(z_{i_k})$  which seems to be remarkably well justified (see our numerical results below). In fact, though this result strictly holds only for slow driving, orders of  $\eta^k$  for  $k \gg 1$  become negligible since  $\eta \in [0, 1]$ , hence, justifying our approximation. Furthermore, it is important to note that for the evaluation of the right hand side of Eq. (28) we only need to solve for the *average* evolution of the system (as dictated by the master equation); it is not necessary to have access to higher order statistics.

*Instantaneous change of driving protocol ("quench").* For a quench we assume that at  $t_m$  with  $0 < t_m < t_f$  the energy levels are shifted instantaneously and are held constant before and after. Then, the MJE is given by (see Appendix A.4 for the derivation)

$$\langle e^{-\beta(W^m - \Delta F)} \rangle_{\mathbf{y}} = 1 - \eta \left[ 1 - p_g^\dagger(t_m) e^{\beta \Delta \omega^\dagger} - p_e^\dagger(t_m) e^{-\beta \Delta \omega^\dagger} \right] \quad (29)$$

where  $\Delta \omega^\dagger \equiv \omega_{\lambda^\dagger(t_f)} - \omega_{\lambda^\dagger(0)}$  and  $\omega_{\lambda^\dagger(t)}$  is defined as before. Note that both relations for the TLS (Eq. (28) and (29)) give the original JE for perfect measurement ( $\eta = 0$ ).



**Figure 3.** Left: Comparison of MC simulation (marks) and numerical integration of the right hand side of Eq. (28) for a continuously driven TLS, where  $\alpha = 1.6\omega_0^2$  and  $\omega_0 \equiv 1$ . Right: Simulation results of a quench of the TLS at  $t_m = t_f/2$  compared with numerical evaluation of Eq. (29) with  $\alpha' = 2.0\omega_0$  and  $\omega_0 \equiv 1$ . For both driving schemes  $\Gamma = 10^{-7}/\Delta t$ ,  $\Delta t = 0.001$  and  $t_f = 3.0$ .

*Numerics.* To test these expressions, we performed Monte Carlo (MC) simulations for different values of  $\eta \in [0, 0.3]$  for two driving schemes. First, the driving scheme varies the energy levels continuously and linearly in time, i.e.,  $\omega_{\lambda(t)} = \omega_0 + \alpha t$ . In Fig. 3 (left) we

plotted the left hand side of Eq. (28) from MC simulations (marks) and the right hand side from numerical integration of the associated ME of the backward protocol (line). As one can see, the approximation of the MJE, Eq. (28), is in very good agreement with the simulation results for small values of  $\eta$ . Note that a value of  $\eta = 0.3$  corresponds to a very large error of the conditional probability  $p_m(y_k|z_k)$  because for a value of  $\eta = 0.5$  the measurement becomes identical to inferring the system state by a fair coin toss. We also test Eq. (29) where we change the driving protocol instantaneously, i.e.,  $\omega_{\lambda(t)} = \omega_0 + \alpha' \Theta(t - t_m)$ . Here, we find perfect agreement of simulation (marks) and numerical integration (line), which is shown in Fig. 3 (right).

#### 4. Measured Jarzynski equality with feedback

Feedback describes the situation in which the state of the system is measured and the evolution of the system is manipulated by applying an external control scheme depending on the measurement outcome. The change of the JE and other fluctuation theorems under feedback has recently attracted a lot of attention, in theory [24–33] as well as in experiments [34,35]. A prominent and the first example of a generalized JE incorporating feedback by performing a single measurement on a stochastic thermodynamic system at a time  $t_m$  with measurement outcome  $y_m$  is the relation derived by Sagawa and Ueda [24]:

$$\langle e^{-\beta(W[\mathbf{z}|y_m] - \Delta F(y_m))} \rangle_{\mathbf{z}, y_m} = \gamma. \quad (30)$$

The so-called efficacy parameter  $\gamma$ , which determines “how efficiently we use the obtained information with feedback control” [24], depends on the probability  $p_{\lambda^\dagger(y_m)}(y_m^*)$  of obtaining the time-reversed outcome  $y_m^*$  in the backward process:

$$\gamma = \int dy_m p_{\lambda^\dagger(y_m)}(y_m^*). \quad (31)$$

Note that in the backward process we use the time-reversed driving protocols  $\lambda^\dagger(t, y_m)$  according to the measurement statistics of  $y_m$  obtained in the forward process. Especially, there is no feedback control in the backwards process.

Now, in the derivation of Eq. (30), the particular measurement yielding outcome  $y_m$  (on which the feedback control is based) is allowed to have measurement errors. However, the left hand side of Eq. (30) is evaluated along the *system* trajectories  $\mathbf{z}$ , which may be inaccessible, especially from an experimental point of view where our knowledge about the situation is solely based on the *measurement* trajectories  $\mathbf{y}$ . We therefore propose a generalization of the JE under feedback control where measurement errors are taken consistently into account. Starting with a general description in Sec. 4.1 we look again at the two specific examples of an OBP in a harmonic potential including a model of an information ratchet in Sec. 4.2 and a feedback controlled TLS in Sec. 4.3 and verify our analytic results by simulations. Furthermore, in Sec. 5 we discuss the relation of the MJE under feedback and the mutual information.

#### 4.1. General case

Let us suppose we measure our system as we did without feedback control but at one instance in time, denoted  $t_m$  with  $0 < t_m < t_f$ , the protocol is changed according to the measurement outcome  $y_m$  such that the protocol is fixed before  $t_m$ , i.e.,  $\lambda = \lambda(t)$  for  $t \in [0, t_m]$  and is dependent on  $y_m$  after  $t_m$ , i.e.,  $\lambda = \lambda(t, y_m)$  for  $t \in (t_m, t_f]$ . The work applied to the system, which now depends on  $y_m$ , is given by

$$W[\mathbf{z}|y_m] = \int_0^{t_m} dt \dot{\lambda}(t) \frac{\partial H_{\lambda(t)}[\mathbf{z}(t)]}{\partial \lambda} + \int_{t_m}^{t_f} dt \dot{\lambda}(t, y_m) \frac{\partial H_{\lambda(t, y_m)}[\mathbf{z}(t)]}{\partial \lambda(y_m)}. \quad (32)$$

The same equation holds also for the measured work  $W_m[\mathbf{y}|y_m]$  by interchanging  $\mathbf{z}$  with  $\mathbf{y}$  (keeping  $y_m$ ). The probability of a path in phase space  $(\mathbf{z}, \mathbf{y})$  under feedback control is denoted by  $\mathcal{P}_{\lambda(y_m)}[\mathbf{z}, \mathbf{y}]$  and we again assume that it factorizes into the probability density of the system trajectory  $\mathcal{P}_{\lambda(y_m)}[\mathbf{z}]$ , which now explicitly depends on  $y_m$ , and the conditional probabilities  $\prod_i p_m(y_i|z_i)$  (see Eq. (16)). Then, the MJE with feedback control can be expressed as

$$\begin{aligned} \langle e^{-\beta(W_m[\mathbf{y}|y_m] - \Delta F(y_m))} \rangle_{\mathbf{y}} &= \int \mathcal{D}[\mathbf{z}] \mathcal{D}[\mathbf{y}] \mathcal{P}_{\lambda(y_m)}[\mathbf{z}, \mathbf{y}] e^{-\beta(W_m[\mathbf{y}|y_m] - \Delta F(y_m))} \\ &= \int \mathcal{D}[\mathbf{z}] \mathcal{D}[\mathbf{y}] \mathcal{P}_{\lambda(y_m)}[\mathbf{z}] \prod_{i=0}^N p_m(y_i|z_i) e^{-\beta(W_m[\mathbf{y}|y_m] - \Delta F(y_m))}. \end{aligned} \quad (33)$$

Note, that the difference in free energy does now also depend on the measurement outcome, i.e.,  $\Delta F = \Delta F(y_m)$ , because the Hamiltonian of the system at time  $t_f$  depends on  $y_m$ . Using again the condition of microreversibility (see Eq. (9)) and assuming time-reversal symmetry of the conditional probabilities,  $p_m(y_i|z_i) = p_m(y_i^*|z_i^*)$ , the following equation holds:

$$\begin{aligned} \langle e^{-\beta(W_m[\mathbf{y}|y_m] - \Delta F(y_m))} \rangle_{\mathbf{y}} &= \int \mathcal{D}[\mathbf{z}^\dagger] \mathcal{D}[\mathbf{y}^\dagger] \mathcal{P}_{\lambda^\dagger(y_m)}[\mathbf{z}^\dagger] \prod_{i=0}^N p_m(y_i^*|z_i^*) e^{-\beta(\Delta e^\dagger(y_m^*) - \delta q^\dagger(y_m^*) + W_m[\mathbf{y}|y_m])} \\ &= \int \mathcal{D}[\mathbf{z}^\dagger] \mathcal{D}[\mathbf{y}^\dagger] \mathcal{P}_{\lambda^\dagger(y_m)}[\mathbf{z}^\dagger, \mathbf{y}^\dagger] e^{-W[\mathbf{z}^\dagger|y_m]} e^{\beta W_m[\mathbf{y}^\dagger|y_m]}. \end{aligned} \quad (34)$$

From Eq. (34) we immediately obtain the MJE in the presence of feedback control:

$$\langle e^{-\beta(W_m[\mathbf{y}|y_m] - \Delta F(y_m))} \rangle_{\mathbf{y}} = \left\langle e^{\beta(W_m^\dagger[\mathbf{y}^\dagger|y_m] - W^\dagger[\mathbf{z}^\dagger|y_m])} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger}, \quad (35)$$

which looks remarkably similar to Eq. (20). Here,  $W_m^\dagger[\mathbf{y}^\dagger|y_m]$  and  $W^\dagger[\mathbf{z}^\dagger|y_m]$  are the measured and true work, respectively, in the backward process applying the time-reversed protocol  $\lambda^\dagger(t, y_m)$  according to the measurement outcome  $y_m$  in the forward process. We stress that we do not perform any feedback in the backward process

equivalently to [24]. Analogously to the efficacy parameter  $\gamma$  (see Eqs. (30) and (31)) we call the right hand side of Eq. (35) measured efficacy parameter,

$$\gamma_m \equiv \left\langle e^{\beta(W_m^\dagger[\mathbf{y}^\dagger|y_m] - W^\dagger[\mathbf{z}^\dagger|y_m])} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger}, \quad (36)$$

because the JE is evaluated using the measured trajectories. Note the subtle distinction between Eq. (30) and (35). Eq. (30) starts with  $\langle \exp(-\beta(W - \Delta F)) \rangle_{\mathbf{z}}$  which experimentally requires an error-free detector to evaluate it. We instead start with  $\langle \exp(-\beta(W - \Delta F)) \rangle_{\mathbf{y}}$  which can be directly evaluated also with a faulty detector. Our final theoretical result (36) then depends on  $\mathbf{z}^\dagger$  indeed. However, based on this definition we show below how to overcome this difficulty for various examples. Furthermore, note that a complementary analytical analysis confirming our results has been reported in Ref. [?] for the example of the Szilard engine.

In the limiting case of perfect measurement,  $p_m(y_k|z_k) = \delta_{y_k, z_k}$ , Eq. (35) simplifies to

$$\begin{aligned} \langle e^{-\beta W_m[\mathbf{y}|y_m] - \Delta F(y_m)} \rangle_{\mathbf{y}} &= \int \mathcal{D}[\mathbf{z}^\dagger] \mathcal{D}[\mathbf{y}^\dagger] \mathcal{P}_{\lambda^\dagger(y_m)}[\mathbf{z}^\dagger] \prod_i \delta_{y_i^*, z_i^*} e^{-\beta(W_m^\dagger[\mathbf{y}^\dagger|y_m] - W^\dagger[\mathbf{z}^\dagger|y_m])} \\ &= \int \mathcal{D}[\mathbf{y}] P_{\lambda^\dagger(y_m)}[\mathbf{y}^\dagger] = \int dy_N^* \dots \int dy_0^* p_{\lambda_N^*(y_m)}(y_N^*) \dots p_{\lambda_0^*}(y_1^* \rightarrow y_0^*). \end{aligned} \quad (37)$$

Due to normalization of conditional probabilities, it holds that the integrals of all  $y_k^*$  with  $k < m$  are equal to unity, hence,

$$\begin{aligned} &\int dy_N^* \dots \int dy_0^* p_{\lambda_N^*(y_m)}(y_N^*) \dots p_{\lambda_0^*}(y_1^* \rightarrow y_0^*) \\ &= \int dy_N^* \dots \int dy_m^* p_{\lambda_N^*(y_m)}(y_N^*) \dots p_{\lambda_m^*}(y_{m+1}^* \rightarrow y_m^*) = \int dy_m^* p_{\lambda_m^*(y_m)}(y_m^*) \\ &= \int dz_m^* p_{\lambda^\dagger(z_m)}(z_m^*) \end{aligned} \quad (38)$$

Only in this case the efficacy  $\gamma$  and the measured efficacy  $\gamma_m$  are the same as it should be.

However, for a measurement outcome  $y_m$  including errors,  $\gamma$  deviates from  $\gamma_m$ . The interpretation and physical significance of the difference between  $\gamma$  and  $\gamma_m$  can be explained as follows: consider two observers Alice and Bob. Suppose that Alice measures the state of the system with a faulty detector whereas Bob measures the system with a perfect detector. Furthermore, suppose that only Alice performs the feedback control based on *her* measurement result at time  $t_m$ . Then, if Alice evaluates the JE of the work done on the system along her measured trajectories, she will observe the result  $\gamma_m$ . In contrast, Bob – given the correct system trajectories and knowledge about the feedback action of Alice and her faulty detector – is able to verify the standard Sagawa-Ueda relation with the efficacy parameter  $\gamma$ .

#### 4.2. Overdamped Brownian motion

As an explicit example, for which we can evaluate the right hand side of Eq. (35) analytically, we look again at an OBP in a harmonic potential (see Sec. 3.2) and assume that the center of the potential is initially at  $\mu_{\lambda(0)} = 0$  and the width is  $f_{\lambda(0)}$ . Both parameters will be changed instantaneously at time  $t_m$  if the measured position at that time is  $y_m > 0$ , the position to  $\mu_{\lambda(t_f)}$  and the stiffness to  $f_{\lambda(t_f)}$ . Otherwise, for  $y_m < 0$ , the potential remains unchanged. For this specific example Eq. (35) can be evaluated explicitly and we obtain (see Appendix A.5)

$$\gamma_m = \frac{1}{2} \left( 1 + \frac{1}{\sqrt{1 + \frac{\kappa}{f_{\lambda(t_f)}} \Delta f}} \operatorname{erfc} \left[ -\mu_{t_f} \sqrt{\frac{\beta f_{\lambda(t_f)} (1 + \kappa)}{1 + \frac{\kappa}{f_{\lambda(t_f)}} \Delta f}} \right] \exp \left\{ \frac{\kappa \beta f_{\lambda(0)} \mu_{t_f}^2}{1 + \frac{\kappa}{f_{\lambda(t_f)}} \Delta f} \right\} \right) \quad (39)$$

where  $\kappa \equiv 2\beta f_{\lambda(0)} \sigma_m^2$ .

For the special case of only altering  $f_{\lambda(t)}$  and keeping the position of the parabola fixed, i.e.,  $\mu_{\lambda(t_f)} = \mu_{\lambda(0)}$ , Eq. (39) reduces to

$$\gamma_m = \frac{1}{2} \left[ 1 + \left( 1 + 2\beta \frac{f_{\lambda(0)}}{f_{\lambda(t_f)}} \Delta f \sigma_m^2 \right)^{-1/2} \right]. \quad (40)$$

On the other hand, if the stiffness is held constant,  $f_{\lambda(t_f)} = f_{\lambda(0)} = f$ , but the parabola is shifted, we find

$$\gamma_m = \frac{1}{2} \left( 1 + e^{2(f\beta\mu_{t_f}\sigma_m)^2} \operatorname{erfc} \left[ -\mu_{t_f} \sqrt{f\beta(1 + 2f\beta\sigma_m^2)} \right] \right). \quad (41)$$

We have verified Eqs. (39) - (41) by performing BD simulations for various driving schemes (not shown here) and will discuss the paradigmatic model of an "information ratchet" [24] in the next paragraph in more detail also showing numerical results.

*Information ratchet.* The Brownian particle is initially in thermal equilibrium in the harmonic potential with center  $\mu_0$ . We then measure the position of the particle  $y_m$  at time  $t_m$  and perform the following feedback scheme: If  $y_m \geq \mu_0 + L$  with  $L > 0$  being constant, we shift the center of the potential  $\mu_{t>t_m} = \mu_0 + 2L$ , if  $y_m < \mu_0 + L$  we do nothing. We then replace  $\mu_0 \rightarrow \mu_0 + 2L$  and start over again after some transient relaxation time. By repeatedly performing this feedback protocol, we can actually move the average position of the particle to the right, ideally without performing work. Here,  $\Delta F = 0$  holds throughout the whole process. Furthermore, one can also extract work from the system by this feedback control if the particle is transported against a potential gradient as, e.g., in the experiment [34]. For a single step of the ratchet, where we put  $\mu_0 = 0$  for simplicity, the measured efficacy with feedback control is given by

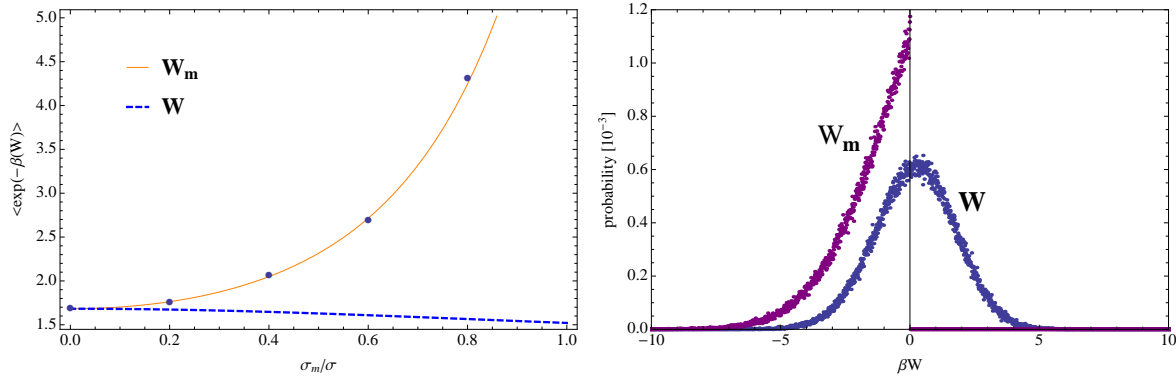
$$\gamma_m = \frac{1}{2} \left( \operatorname{erfc} \left[ -\frac{L}{\sqrt{\frac{1}{\beta f} + 2\sigma_m^2}} \right] + e^{8f^2 L^2 \beta^2 \sigma_m^2} \operatorname{erfc} \left[ -L \frac{1 + 4f\beta\sigma_m^2}{\sqrt{\frac{1}{\beta f} + 2\sigma_m^2}} \right] \right). \quad (42)$$



The derivation follows the same steps as in Appendix A.5 but the integral of  $y_m^*$  is splitted at  $L$  instead of 0. Eq. (42) differs from the efficacy parameter  $\gamma$  of the original information ratchet [24],

$$\gamma = \text{erfc} \left[ -\frac{L}{\sqrt{\frac{1}{\beta f} + 2\sigma_m^2}} \right]. \quad (43)$$

In Fig. 4 (left) we plot the solutions of the two equations above as function of the variance of the measurement  $\sigma_m$ . The two equations coincide for the case of perfect



**Figure 4.** Left: efficacy parameter  $\gamma$  (dashed) and measured efficacy  $\gamma_m$  (line) as function of the measurement error as well as results from BD simulation (marks) with  $f = 2.0$  and  $L = 0.5$ . Right: Probability distribution of the work extracted and performed by the system (blue) and the extracted measured work (purple) for the information ratchet with  $\sigma_m/\sigma = 1.0$  where  $\sigma = (2\beta f)^{-1/2}$ .

measurement. However, for finite values of  $\sigma_m$  the efficacy  $\gamma$  of the feedback control (dashed line) is lower than for perfect measurement: If the measurement has an error, then the potential will be shifted even though the real position of the particle may not be greater than  $L$ . Then we may actually apply work to the system instead of extracting it and the average value of extracted work is lower for noisier measurements.

If we look at the work we measure using the same apparatus as we have used to measure  $y_m$  (line), we see that with increasing measurement error  $\sigma_m$ , the measured efficacy  $\gamma_m$  also increases in strong contrast to  $\gamma$ . Since the measured work is given in terms of the measured position  $y_m$  of the particle, we always apply the “correct” feedback scheme from the observer’s point of view. Thus, we (the observer) always think that we extract work. This can also be seen in the distribution of measured (purple) and system (blue) work in Fig. 4 (right), where the probability of measured work is only non-zero for  $W_m < 0$ . To support this claim even further, we can calculate the average measured and system work by integration of Eq. (32) over  $z$  and  $y_m$ , where the integral is nonzero only if  $y_m > L$ . The difference of them results in

$$\langle W[y|y_m] \rangle - \langle W[z|y_m] \rangle = -4fL\sigma_m^2 \sqrt{\frac{f\beta}{\pi(1+\kappa)}} \exp \left\{ -\frac{f\beta L^2}{1+\kappa} \right\} \leq 0 \quad (44)$$

where  $\kappa = 2\beta f\sigma_m^2$ . Thus, on average the measured *extracted* work (note that in our convention work is positive if it is done on the system) from the system will be greater than the true extracted work and even increases with  $\sigma_m$ . For a larger value of  $\sigma_m$  the probability distribution  $p'_m(y_m)$  (see Eq. (13)) of the measured position  $y_m$  is broader (i.e., has a larger variance) than  $p(z)$ , but still has the same mean value as  $p(z)$ . Then, measurement outcomes with  $y_m > L$  are more frequent and  $\gamma_m$  increases.

#### 4.3. Two-level system

Similarly to the derivation of the MJE of the TLS without feedback we find with feedback for a c.p.d. but at this point unspecified driving protocol an approximation for the modification of the original JE (see Appendix A.6 for details):

$$\gamma_m \approx \sum_{z_m \in \{g, e\}} \left[ (1 - \eta) p_{\lambda^\dagger(z_m)}(z_m) \exp \left( -\eta\beta \int dt \dot{\omega}_{\lambda^\dagger(t, z_m)} (p_{e, \lambda^\dagger(z_m)}(t) - p_{g, \lambda^\dagger(z_m)}(t)) \right) \right. \\ \left. - \eta p_{\lambda^\dagger(\bar{z}_m)}(z_m) \exp \left( -\eta\beta \int dt \dot{\omega}_{\lambda^\dagger(t, \bar{z}_m)} (p_{e, \lambda^\dagger(\bar{z}_m)}(t) - p_{g, \lambda^\dagger(\bar{z}_m)}(t)) \right) \right] \quad (45)$$

Here,  $p_{z, \lambda^\dagger(y_m)}(t)$  is the probability for the system to be in state  $z$  (ground or excited) at time  $t$  in the backward process with the backward protocol according to the measurement outcome  $y_m$  (ground or excited state) in the forward process. We again note that we do not apply feedback in the backward process and that Eq. (45) is valid under exactly the same conditions as discussed below Eq. (28). Furthermore,  $\omega_{\lambda^\dagger(t, y_m)}$  is the energy gap as defined in Sec. 3.3 with the time-reversed protocol according to the outcome of the forward process. For a c.p.d. protocol with nondifferentiable points the integral in Eq. (45) is again split into parts at the respective points. For most driving protocols with feedback we have considered numerically (not shown here) Eq. (45) is a very good approximation.

For a driving protocol that is not continuous in time, we find a different expression. Here, we assume as in the case without feedback, that before and after  $t_m$  the protocol is constant and that a quench is performed at time  $t_m$ . We then find for the MJE (see also Appendix A.6)

$$\gamma_m = \sum_{z_m \in \{g, e\}} \left[ (1 - \eta) p_{z_m, \lambda^\dagger(z_m)}(t_m) + \eta p_{z_m, \lambda^\dagger(\bar{z}_m)}(t_m) e^{\beta \Delta \omega_{\lambda^\dagger(\bar{z}_m)}(z_m)} \right] \quad (46)$$

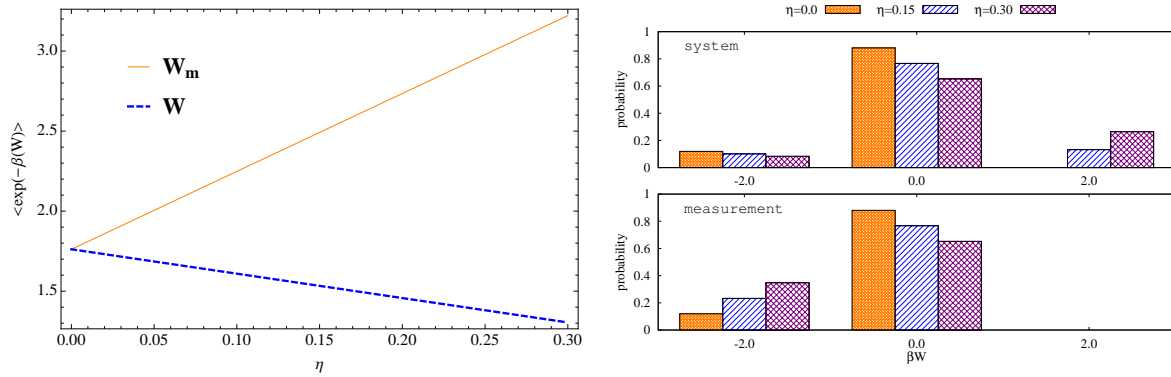
where  $p_{z_m, \lambda^\dagger(\bar{z}_m)}(t_m)$  denotes the probability of the system to be in state  $z_m$  at time  $t_m$  in the backward process with the backward protocol according to the measurement outcome  $y_m = \bar{z}_m$ . Here, we introduced the complementary state  $\bar{z}_m$  to  $z_m$  (i.e., if  $z_m = g$  then  $\bar{z}_m = e$  and vice versa). Furthermore,  $\Delta \omega_{\lambda^\dagger(\bar{z}_m)}(z_m) \equiv \omega_{\lambda^\dagger(t_f, \bar{z}_m)}(z_m) - \omega_{\lambda^\dagger(0, \bar{z}_m)}(z_m)$  and  $\omega_{\lambda^\dagger(t, \bar{z}_m)}(z_m) = \varepsilon_{\lambda^\dagger(t, \bar{z}_m)}(z_m) - \varepsilon_{\lambda^\dagger(t, \bar{z}_m)}(\bar{z}_m)$ .

We will now discuss an example of a protocol with a quench in detail in the next paragraph.

*Conditional swap.* As a specific example, for which we can extract work from a single heat bath by measuring the state of the TLS at time  $t_m$ , we discuss a feedback operation which we call a conditional swap: if at time  $t_m$  the measured state of the TLS  $y_m$  is the excited state, we interchange the two energy levels such that we extract work of  $\omega = \varepsilon_e - \varepsilon_g$  if the system state  $z_m$  is the excited one and perform work of  $-\omega$  if  $z_m = g$ . If  $y_m = g$  we do nothing. We compare our findings (see Eq. (46)) of this conditional swap to the corresponding expression of the efficacy parameter  $\gamma$ , which is given for this specific example by

$$\gamma = (1 - \eta)2p_{g,\lambda^\dagger(g)}(t_m) + 2\eta p_{e,\lambda(g)}(t_m) . \quad (47)$$

Note that in the model of the conditional swap  $p_{g,\lambda^\dagger(g)}(t_m) = p_{e,\lambda^\dagger(e)}(t_m)$  and  $p_{e,\lambda^\dagger(g)}(t_m) = p_{g,\lambda^\dagger(e)}(t_m)$ .



**Figure 5.** Left: Efficacy parameter  $\gamma$  of the system (dashed) and measured efficacy  $\gamma_m$  (line) as function of the measurement error  $\eta$  as well as results from MC simulation (marks) for the conditional swap operation at  $t_m = t_f/2$  with  $\Gamma = 10^{-7}/\Delta t$ ,  $\Delta t = 0.001$  and  $t_f = 3.0$ . Right: Work distribution of the system (top) and the measured work distribution (bottom) of the conditional swap for different values of  $\eta$ .

We show the difference of  $\gamma$  (dashed) and  $\gamma_m$  (line) for different values of  $\eta$  in Fig. 5 (left). As one can see, for a perfect measurement they result in the same value. However if  $\eta$  is greater than zero, the two differ. The explanation is very similar to the one of the information ratchet discussed in Sec. 4.2: if the measurement  $y_m$  involves errors, the two states are sometimes interchanged even though the system may be in the ground state resulting in work *applied to* the system instead of extracting work from the system. If we look at the work distribution of the system (see Fig. 5 right top), one can see that for values  $\eta > 0$ , the extracted work becomes less whereas the probability of applying work to the system increases with measurement error (note that in our convention work is negative if it is done by the system). Then the efficacy parameter is lower than without measurement error. On the other hand, if we look at the measured work (see Fig. 5 right bottom), which is calculated from the measured state of the system, we only measure positive work extraction from the system by performing the conditional swap. Furthermore, the probability of measuring the excited state of the system is

always larger than the actual probability of the system to be in the excited state if  $p_e(t_m) < 1/2$  (as in our case),

$$p'_{y_m=e}(t_m) = (1 - \eta)p_e(t_m) + \eta p_g(t_m) = p_e(t_m) + \eta(1 - 2p_e(t_m)) \geq p_e(t_m) . \quad (48)$$

Therefore, the probability of extracting work from the system and therefore  $\gamma_m$  increases with larger values of  $\eta$ .

## 5. Jarzynski equality with mutual information

We have seen that the classic JE  $\langle e^{-\beta(W-\Delta F)} \rangle = 1$  in general holds only if the system is observed perfectly and no feedback is performed. If one of the conditions is violated, we have in general  $\langle e^{-\beta(W-\Delta F)} \rangle \neq 1$ . However, in case of feedback at a given time  $t_m$  Sagawa and Ueda and others have found that [24–33]

$$\langle e^{-\beta(W[\mathbf{z}|y_m]-\Delta F(y_m))-I(z_m, y_m)} \rangle_{\mathbf{z}, y_m} = 1. \quad (49)$$

Thus, by adding the stochastic mutual information  $I(z_m, y_m) \equiv \ln \frac{p(y_m, z_m)}{p(y_m)p(z_m)}$  to the exponent we can make the right hand side of the “Jarzynski-Sagawa-Ueda relation” equal to unity again. This result provides us with a nice interpretation because it tells us that the amount of work we can extract from the system is bounded by  $\langle I(y_m, z_m) \rangle_{z_m, y_m}$ , which can be viewed as the amount of correlations established during the measurement.

Unfortunately, in case of measurement errors, validating Eq. (49) requires to be able to observe the system *perfectly* during the time where it is not controlled. But this again raises the question of how this might be achieved because this means that the detector of the experimentalist is only faulty previous to the feedback step and otherwise correct. Eq. (49) could be therefore viewed as an “objective” fluctuation theorem which a second “superobserver” with perfect access to both the system *and* detector degrees of freedom would observe. In contrast, the MJE we have considered so far could be called a “subjective” fluctuation theorem which is based on the knowledge of the observer only.

In fact, we will now show that by taking the full stochastic mutual information between the system and detector into account, defined as

$$I[\mathbf{z}, \mathbf{y}] = \ln \left( \frac{\mathcal{P}[\mathbf{z}, \mathbf{y}]}{\mathcal{P}[\mathbf{z}]P[\mathbf{y}]} \right), \quad (50)$$

yields a fluctuation theorem of the form

$$\langle e^{-\beta(W[\mathbf{z}|\mathbf{y}]-\Delta F(\mathbf{y}))-I(\mathbf{z}, \mathbf{y})} \rangle_{\mathbf{z}, \mathbf{y}} = 1 \quad (51)$$

which holds without and with measurement errors and without and with feedback, even if the feedback is performed continuously, i.e., every time step  $\delta t$ . However, the latter relation may be invalid for some error-free feedback control processes where absolute irreversibility is inherent [?]. We remark that the validity of Eq. (51) without feedback and with measurement errors was already noted in Ref. [13] and with feedback with or without measurement errors in Refs. [26, 28, 31]

To prove Eq. (51) we note the chain of equalities

$$\begin{aligned}
\langle e^{-\beta(W[\mathbf{z}|\mathbf{y}] - \Delta F(\mathbf{y})) - I(\mathbf{z}, \mathbf{y})} \rangle_{\mathbf{z}, \mathbf{y}} &= \int \mathcal{D}[\mathbf{y}] \mathcal{D}[\mathbf{z}] \mathcal{P}[\mathbf{z}, \mathbf{y}] e^{-\beta(W_m[\mathbf{z}|\mathbf{y}] - \Delta F(\mathbf{y}))} \frac{\mathcal{P}[\mathbf{z}] \mathcal{P}[\mathbf{y}]}{\mathcal{P}[\mathbf{z}, \mathbf{y}]} \\
&= \int \mathcal{D}[\mathbf{y}] \mathcal{P}[\mathbf{y}] \int \mathcal{D}[\mathbf{z}] \mathcal{P}[\mathbf{z}] e^{-\beta(W_m[\mathbf{z}|\mathbf{y}] - \Delta F(\mathbf{y}))} \\
&= \int \mathcal{D}[\mathbf{y}] \mathcal{P}[\mathbf{y}] = 1.
\end{aligned} \tag{52}$$

Here, we used that the JE  $\int \mathcal{D}[\mathbf{z}] \mathcal{P}[\mathbf{z}] e^{-\beta(W_m[\mathbf{z}|\mathbf{y}] - \Delta F(\mathbf{y}))} = 1$  holds for every fixed measurement record  $\mathbf{y}$  and (consequently in case of feedback) any control protocol  $\lambda(t, \mathbf{y})$ .

Thus, the mutual information seems to be a universal quantity in order to establish fluctuation theorems where not only the system but also the detector has to be taken into account, although it does not possess an obvious thermodynamic interpretation in case without feedback. Unfortunately, finding some (non-trivial) quantity  $G = G[\mathbf{y}]$  such that the MJE can be corrected, i.e., such that  $\langle e^{-\beta(W - \Delta F) - G} \rangle_{\mathbf{y}} = 1$ , remains an open problem at the moment.

## 6. Conclusions and Outlook

In the present paper, we generalized the original JE expressed in terms of the "true" work done on the system to an equation for arbitrary measurement errors based on the measurement record  $\mathbf{y}$ . The key ingredient for this was the conditional probability distribution  $p_m(y|z)$ , which quantifies the uncertainty of a measurement outcome  $y$  given that the system state is  $z$  and which defines an abstract measurement model. In fact, by shifting the attention from  $\mathbf{z}$  to  $\mathbf{y}$  we only did a first step in generalizing stochastic thermodynamics to the presence of measurement errors because much more sophisticated inference schemes could have been considered as well (we actually did not even use Eq. (14) in our derivations leaving this interesting problem to future work).

Then, using the formalism of stochastic path integrals, we derived the MJE (measured JE) without feedback (Eq. (20)) and with feedback control (Eq. (35)). These expressions were general (under the assumption of a Markovian measurement apparatus), but explicitly involve system trajectory dependent quantities. For two important paradigmatic examples we could overcome this difficulty and express the MJE in terms of fixed Hamiltonian parameters or average quantities, which can be computed based on a master equation. For an OBP trapped in a harmonic potential the expressions derived were exact, whereas for the TLS exact solutions were only found for quenches and very good approximations for continuous driving protocols. We also checked our findings with simulation results. In the limiting case of perfect measurement the general MJE equations result in the original JE without and with feedback. For the non-ideal case we hope that our theory provides a convenient way to explain the always noisy statistics in experiments, which have beautifully demonstrated the validity of the

JE and other fluctuation theorems within the given statistical accuracy so far, see, e.g., Refs. [36–44].

Furthermore, in case of feedback control the correct handling of measurement errors is even more important because we put the obtained information back into the system to influence its future behaviour. Here, we have seen that the measured efficacy  $\gamma_m$  may exceed the system efficacy  $\gamma$  and, contrary to previous intuition, increases with larger measurement errors, which we have calculated explicitly for an information ratchet of an OBP and a conditional swap of the TLS. Furthermore, we showed that the "Jarzynski-Sagawa-Ueda relation" by incorporating the full stochastic mutual information always holds for a "superobserver" who has access to the measured *and* system trajectories, without and with measurement errors and without and with feedback.

Finally, we would like to mention that a lot of research has already been carried out to understand the stochastic thermodynamics of coarse-grained systems, see, e.g., Refs. [45–55]. In there, given a set of microstates, a subset of observable states is introduced, which defines the coarse-graining and which is sometimes explicitly modeled by a detector or sensor. Based on the observability of this subset, the changed laws of (stochastic) thermodynamics are investigated. Though one can argue that both approaches pursue the same research goal, it is worthwhile to point out that our approach is in principle different. First, the coarse-graining approach still assumes that it is possible to observe the particular subsets perfectly, i.e., error-free, and second, it is also implicitly assumed that it is actually possible to find these subsets or to physically model a detector, but this might be challenging for some large detectors such as a camera. Nevertheless, the question to what extent our approach based on an abstract measurement model  $p_m(y|z)$  is equivalent to an explicit detector model with underlying coarse-grained system dynamics is, in our point of view, interesting to study in the future.

## Acknowledgments

Financial support of the DFG through project GRK 1558 is gratefully acknowledged.

## References

- [1] M. Esposito, U. Harbola, and S. Mukamel. Nonequilibrium fluctuations, fluctuation theorems, and counting statistics in quantum systems. *Rev. Mod. Phys.*, 81:1665, 2009.
- [2] K. Sekimoto. *Stochastic Energetics*. Lect. Notes Phys., Springer, Berlin Heidelberg, 2010.
- [3] M. Campisi, P. Hänggi, and P. Talkner. Colloquium: Quantum fluctuation relations: Foundations and applications. *Rev. Mod. Phys.*, 83:771, 2011.
- [4] C. Jarzynski. Equalities and inequalities: irreversibility and the second law of thermodynamics at the nanoscale. *Annu. Rev. Condens. Matter Phys.*, 2:329–351, 2011.
- [5] U. Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep. Prog. Phys.*, 75:126001, 2012.
- [6] G. Schaller. *Open Quantum Systems Far from Equilibrium*. Lect. Notes Phys., Springer, Cham, 2014.



- [7] C. Van den Broeck and M. Esposito. Ensemble and trajectory thermodynamics: A brief introduction. *Physica (Amsterdam)*, 418A:6–16, 2015.
- [8] M. Ribezzi-Crivellari and F. Ritort. Free-energy inference from partial work measurements in small systems. *Proc. Natl. Acad. Sci.*, 111:3386–3394, 2014.
- [9] A. Alemany, M. Ribezzi-Crivellari, and F. Ritort. From free energy measurements to thermodynamic inference in nonequilibrium small systems. *New Journal of Physics*, 17:075009, 2015.
- [10] J. Bechhoefer. Hidden markov models for stochastic thermodynamics. *New. J. Phys.*, 17:075003, 2015.
- [11] K. L. Viisanen, S. Suomela, S. Gasparinetti, O.-P. Saira, J. Ankerhold, and J. P. Pekola. Incomplete measurement of work in a dissipative two level system. *New. J. Phys.*, 17:055014, 2015.
- [12] J. J. Alonso, E. Lutz, and R. Alessandro. Thermodynamics of weakly measured quantum systems. *Phys. Rev. Lett.*, 116:080403, 2016.
- [13] R. García-García, L. Sourabh, and D. Lacoste. Thermodynamic inference based on coarse-grained data or noisy measurements. *Phys. Rev. E*, 93:032103, 2016.
- [14] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78:2690, 1997.
- [15] C. Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Phys. Rev. E*, 56:5018–5035, 1997.
- [16] M. Chaichian and A. Demichev. *Path Integrals in Physics: Volume II Quantum Field Theory, Statistical Physics and other Modern Applications*. Institute of Physics, London, 2001.
- [17] G. E. Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems. *J. Stat. Phys.*, 90:1481–1487, 1998.
- [18] G. E. Crooks. Path-ensemble averages in systems driven far from equilibrium. *Phys. Rev. E*, 61:2361–2366, 2000.
- [19] C. Jarzynski. Hamiltonian derivation of a detailed fluctuation theorem. *J. Stat. Phys.*, 98:77–102, 2000.
- [20] G. J. Milburn. Classical and quantum conditional statistical dynamics. *Quantum Semiclass. Opt.*, 8:269, 1996.
- [21] P. Strasberg, G. Schaller, and T. Brandes. Controlling the stability of steady states in continuous variable quantum systems. in *Control of Self-Organizing Nonlinear Systems (Springer International Publishing)*, pages 289–313, 2016.
- [22] G. A. Huber and S. Kim. Weighted-ensemble brownian dynamics simulations for protein association reactions. *Biophys. J.*, 70:97, 1996.
- [23] B. W. Zhang, D. Jasnow, and D. M. Zuckerman. The ”weighted ensemble” path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *J. Chem. Phys.*, 132:054107, 2010.
- [24] T. Sagawa and M. Ueda. Generalized Jarzynski equality under nonequilibrium feedback control. *Phys. Rev. Lett.*, 104:090602, 2010.
- [25] M. Ponnurugan. Generalized detailed fluctuation theorem under nonequilibrium feedback control. *Phys. Rev. E*, 82:031129, 2010.
- [26] J. M. Horowitz and S. Vaikuntanathan. Nonequilibrium detailed fluctuation theorem for repeated discrete feedback. *Phys. Rev. E*, 82:061120, 2010.
- [27] A. Kundu. Nonequilibrium fluctuation theorem for systems under discrete and continuous feedback control. *Phys. Rev. E*, 86:021107, 2012.
- [28] S. Lahiri, S. Rana, and A. M. Jayannavar. Fluctuation theorems in the presence of information gain and feedback. *J. Phys. A: Math. Theor.*, 45:065002, 2012.
- [29] D. Abreu and U. Seifert. Thermodynamics of genuine nonequilibrium states under feedback control. *Phys. Rev. Lett.*, 108:030601, 2012.
- [30] T. Sagawa and M. Ueda. Fluctuation theorem with information exchange: Role of correlations in stochastic thermodynamics. *Phys. Rev. Lett.*, 109:180602, 2012.



- [31] T. Sagawa and M. Ueda. Nonequilibrium thermodynamics of feedback control. *Phys. Rev. E*, 85:021104, 2012.
- [32] K. Funo, Y. Watanabe, and M. Ueda. Integral quantum fluctuation theorems under measurement and feedback control. *Phys. Rev. E*, 88:052121, 2013.
- [33] T. Munakata and M. L. Rosinberg. Entropy production and fluctuation theorems for langevin processes under continuous non-Markovian feedback control. *Phys. Rev. Lett.*, 112:180601, 2014.
- [34] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano. Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. *Nat. Phys.*, 6:988–992, 2010.
- [35] J. V. Koski, V. F. Maisi, T. Sagawa, and J. P. Pekola. Experimental observation of the role of mutual information in the nonequilibrium dynamics of a Maxwell demon. *Phys. Rev. Lett.*, 113:030601, 2014.
- [36] G. Hummer and A. Szabo. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proc. Natl. Acad. Sci.*, 98:3658–3661, 2001.
- [37] G. M. Wang, E. M. Sevick, E. Mittag, D. J. Searles, and D. J. Evans. Experimental demonstration of violations of the second law of thermodynamics for small systems and short time scales. *Phys. Rev. Lett.*, 89:050601, 2002.
- [38] J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, and C. Bustamante. Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski’s equality. *Science*, 296:1832–1835, 2002.
- [39] E. H. Trepagnier, Jarzynski, F. Ritort, G. E. Crooks, C. J. Bustamante, and J. Liphardt. Experimental test of Hatano and Sasa’s nonequilibrium steady-state equality. *Proc. Natl. Acad. Sci.*, 101:15038–15041, 2004.
- [40] S. Schuler, T. Speck, C. Tietz, J. Wrachtrup, and U. Seifert. Experimental test of the fluctuation theorem for a driven two-level system with time-dependent rates. *Phys. Rev. Lett.*, 94:180602, 2005.
- [41] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, and C. Bustamante. Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature (London)*, 437:231–234, 2005.
- [42] Y. Utsumi, D. S. Golubev, M. Marthaler, K. Saito, T. Fujisawa, and G. Schön. Bidirectional single-electron counting and the fluctuation theorem. *Phys. Rev. B*, 81:125331, 2010.
- [43] B. Küng, C. Rössler, M. Beck, M. Marthaler, D. S. Golubev, Y. Utsumi, T. Ihn, and K. Ensslin. Irreversibility on the level of single-electron tunneling. *Phys. Rev. X*, 2:011001, 2012.
- [44] S. An, J. N. Zhang, M. Um, D. Lv, Y. Lu, J. Zhang, Z.-Q. Yin, H. T. Quan, and K. Kim. Experimental test of the quantum Jarzynski equality with a trapped-ion system. *Nat. Phys.*, 11:193–199, 2015.
- [45] A. Puglisi, S. Pigolotti, L. Rondoni, and A. Vulpiani. Entropy production and coarse graining in markov processes. *J. Stat. Mech.*, P05015, 2010.
- [46] G. Bulnes Cuetara, M. Esposito, and P. Gaspard. Fluctuation theorems for capacitively coupled electronic currents. *Phys. Rev. B*, 84:165114, 2011.
- [47] B. Altaner and J. Vollmer. Fluctuation-preserving coarse graining for biochemical systems. *Phys. Rev. Lett.*, 108:228101, 2012.
- [48] J. Mehl, B. Lander, C. Bechinger, V. Blickle, and U. Seifert. Role of hidden slow degrees of freedom in the fluctuation theorem. *Phys. Rev. Lett.*, 108:220601, 2012.
- [49] M. Esposito. Stochastic thermodynamics under coarse graining. *Phys. Rev. E*, 85:041125, 2012.
- [50] P. Strasberg, G. Schaller, T. Brandes, and M. Esposito. Thermodynamics of a physical model implementing a maxwell demon. *Phys. Rev. Lett.*, 110:040601, 2013.
- [51] G. Bulnes Cuetara, M. Esposito, G. Schaller, and P. Gaspard. Effective fluctuation theorems for electron transport in a double quantum dot coupled to a quantum point contact. *Phys. Rev. B*, 88:115134, 2013.
- [52] T. Leonard, B. Lander, U. Seifert, and T. Speck. Stochastic thermodynamics of fluctuating

- density fields: non-equilibrium free energy differences under coarse-graining. *J. Chem. Phys.*, 139:204109, 2013.
- [53] A. C. Barato, D. Hartich, and U. Seifert. Rate of mutual information between coarse-grained non-markovian variables. *J. Stat. Phys.*, 153:460–478, 2013.
- [54] E. Zimmermann and U. Seifert. Effective rates from thermodynamically consistent coarse-graining of models for molecular motors with probe particles. *Phys. Rev. E*, 91:022709, 2015.
- [55] M. Esposito and J. M. R. Parrondo. Stochastic thermodynamics of hidden pumps. *Phys. Rev. E*, 91:052114, 2015.

## Appendix A. Appendix

### Appendix A.1. Derivation of MJE for continuous driving of OBP

In this section we derive the analytic expression of the MJE for an OBP in a harmonic potential in one dimension, namely Eq. (25). We assume the external control parameter  $\lambda(t)$  to be c.p.d. (continuous and piecewise differentiable) throughout this section. The discretized work along a trajectory  $\mathbf{z}$  given the Hamiltonian in Eq. (22) becomes

$$W[\mathbf{z}] = \sum_i (\delta f_{\lambda_i} z_{i-1}^2 - 2\delta[f\mu]_{\lambda_i} z_{i-1} + \delta[f\mu^2]_{\lambda_i}) \quad (\text{A.1})$$

where  $\delta f_{\lambda_i} = f_{\lambda_i} - f_{\lambda_{i-1}}$ ,  $\delta[f\mu]_{\lambda_i} = f_{\lambda_i}\mu_{\lambda_i} - f_{\lambda_{i-1}}\mu_{\lambda_{i-1}}$  and  $\delta[f\mu^2]_{\lambda_i} = f_{\lambda_i}\mu_{\lambda_i}^2 - f_{\lambda_{i-1}}\mu_{\lambda_{i-1}}^2$ . For the example considered here, it holds that  $z_i^* = z_i$  and  $y_i^* = y_i$ .

By factorizing  $\mathcal{P}[\mathbf{z}, \mathbf{y}]$  (see Eq. (16)) one can express the right hand side of the general Eq. (20) as

$$\begin{aligned} \left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} &= \int \mathcal{D}[\mathbf{z}^\dagger] \mathcal{P}^\dagger[\mathbf{z}^\dagger] \int dy_0 \dots \int dy_N \prod_i \left[ p_m(y_i | z_i) \right. \\ &\quad \times \exp \left\{ \beta \left( \delta f_{\lambda_{i+1}^\dagger} y_i^2 - 2\delta[f\mu]_{\lambda_{i+1}^\dagger} y_i - \delta f_{\lambda_{i+1}^\dagger} z_i^2 + 2\delta[f\mu]_{\lambda_{i+1}^\dagger} z_i \right) \right\} \Big] . \end{aligned} \quad (\text{A.2})$$

Assuming a normal distribution of  $p_m(y_i | z_i)$  (see Eq. (24)) we find after integration over all  $y_k$ :

$$\begin{aligned} \left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} &= \left( \prod_i \frac{1}{\sqrt{1 - 2\beta\delta f_{\lambda_{i+1}^\dagger} \sigma_m^2}} \right) \times \\ &\quad \int \mathcal{D}[\mathbf{z}^\dagger] \mathcal{P}^\dagger[\mathbf{z}^\dagger] \exp \left\{ - \sum_i \frac{2\beta^2 \sigma_m^2}{2\beta\delta f_{\lambda_{i+1}^\dagger} \sigma_m^2 - 1} \left( \delta[f\mu]_{\lambda_{i+1}^\dagger} - \delta f_{\lambda_{i+1}^\dagger} z_i \right)^2 \right\} . \end{aligned} \quad (\text{A.3})$$

Note that for the integral over  $y_k$  to converge the standard deviation of the measurement must obey

$$\sigma_m^2 < \frac{1}{2\beta \left| \delta f_{\lambda_{k+1}^\dagger} \right|} . \quad (\text{A.4})$$

This means that in an experimental setup (or also for simulations), in which the width of the potential is varied between two measurements by a finite value  $\delta f_{\lambda_{k+1}^\dagger}$ , the deviation of measured and system coordinate cannot be arbitrarily large.

We first look at the integral of Eq. (A.3): in the limit  $N \rightarrow \infty$  the time steps  $dt = t_f/N$  become infinitesimal and we can write the term in the exponential approximately as

$$\begin{aligned} & \exp \left\{ - \sum_i \frac{2\beta^2 \sigma_m^2}{2\beta \delta f_{\lambda_{i+1}^\dagger} \sigma_m^2 - 1} \left( \delta[f\mu]_{\lambda_{i+1}^\dagger} - \delta f_{\lambda_{i+1}^\dagger} z_i \right)^2 \right\} \\ & \approx \exp \left\{ 2\beta^2 \sigma_m^2 dt \int_0^{t_f} dt \left( [f\mu]'_{\lambda^\dagger(t)} - f'_{\lambda^\dagger(t)} z(t) \right)^2 \right\} \equiv \star \end{aligned} \quad (\text{A.5})$$

where the prime (e.g.,  $f'$ ) denotes a derivative with respect to time  $t$ . Note that the additional  $dt$  in front of the integral is correct. Furthermore, this step is only exact provided that the protocol is differentiable. However, as long as it is continuous and only nondifferentiable at a finite number of points  $0 < t_1 < \dots < t_K < t_f$  this argument can be easily generalized by splitting the integral at the respective places (i.e.,  $\int_0^{t_1} dt + \int_{t_1}^{t_2} dt + \dots + \int_{t_K}^{t_f} dt$ ) and by observing that due to the continuity  $\delta f_{\lambda_{i+1}^\dagger}$  and  $\delta[f\mu]_{\lambda_{i+1}^\dagger}$  remain infinitesimal small at all points. Then, by the mean value theorem of integration we know that there exists a  $\xi \in [0, t_f]$  such that

$$\star = \exp \left\{ 2\beta^2 \sigma_m^2 dt \left( [f\mu]'_{\lambda^\dagger(\xi)} - f'_{\lambda^\dagger(\xi)} z(\xi) \right)^2 \right\}. \quad (\text{A.6})$$

and hence, this term becomes 1 for  $N \rightarrow \infty$ , i.e.,  $dt \rightarrow 0$ .

Therefore, Eq. (A.3) simplifies to

$$\left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} = \prod_i \frac{1}{\sqrt{1 - 2\beta \delta f_{\lambda_{i+1}^\dagger} \sigma_m^2}} \approx \prod_i \left( 1 + \sigma_m^2 \beta \delta f_{\lambda_{i+1}^\dagger} \right), \quad (\text{A.7})$$

which holds for  $N \rightarrow \infty$ . In the last step, we write the product as an exponential and use an approximation of the logarithm up to first order:

$$\left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} = \exp \left( \sum_i \ln \left( 1 + \sigma_m^2 \beta \delta f_{\lambda_{i+1}^\dagger} \right) \right) \approx \exp \left( \sigma_m^2 \beta (f_{\lambda_0} - f_{\lambda_N}) \right). \quad (\text{A.8})$$

Taking the limit  $N \rightarrow \infty$ ,  $f_{\lambda_0} = f_{\lambda(t_0)}$  and  $f_{\lambda_N} = f_{\lambda(t_N)}$ , we arrive at Eq. (25).

## Appendix A.2. Derivation of MJE for instantaneous driving of OBP

Here, we derive Eq. (26), where we assume that the stiffness of the harmonic potential as well as the position are instantaneously changed at the same time  $t_m$ . Since the driving protocol is constant before and after  $t_m$ , it holds that  $\delta f_{\lambda_{k+1}^\dagger} = 0$  as well as  $\delta[f\mu]_{\lambda_{k+1}^\dagger} = 0$  for all  $k \neq m$ . In this case the right hand side of Eq. (20) reads after integration over all  $y_k$  and  $z_k$  with  $k \neq m$ :

$$\begin{aligned} \left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} &= \int dz_m p^\dagger(z_m) \int dy_m \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp \left\{ - \frac{(z_m - y_m)^2}{2\sigma_m^2} \right\} \\ &\quad \times \exp \left\{ \beta \delta f_{\lambda_{m+1}^\dagger} y_m^2 - 2\delta[f\mu]_{\lambda_{m+1}^\dagger} y_m - \delta f_{\lambda_{m+1}^\dagger} z_m^2 + 2\delta[f\mu]_{\lambda_{m+1}^\dagger} z_m \right\}. \end{aligned} \quad (\text{A.9})$$

For a quench it holds that  $\delta f_{\lambda_{m+1}^\dagger} = f_{\lambda(0)} - f_{\lambda(t_f)} \equiv -\Delta f$  and equivalently  $\delta[f\mu]_{\lambda_{m+1}^\dagger} = f_{\lambda(0)}\mu_{\lambda(0)} - f_{\lambda(t_f)}\mu_{\lambda(t_f)} \equiv -\Delta[f\mu]$ . Then, the integration over  $y_m$  yields

$$\begin{aligned} & \left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} \\ &= \int dz_m p^\dagger(z_m) \frac{1}{\sqrt{1 + 2\Delta f \beta \sigma_m^2}} \exp \left\{ \frac{2\beta^2 \sigma_m^2}{1 + 2\Delta f \beta \sigma_m^2} (-\Delta[f\mu] + \Delta f z_m)^2 \right\}. \end{aligned} \quad (\text{A.10})$$

For the integral over  $y_m$  to converge, it must again hold that  $\sigma_m^2 < (2\beta |\Delta f|)^{-1}$ .

We now use, that for the harmonic potential the probability distribution of the position of the OBP in equilibrium (initial system state) is Gaussian distributed with mean  $\mu_{\lambda(t_f)}$  and variance  $(2\beta f_{\lambda(t_f)})^{-1/2}$  in the time-reversed protocol. The integration over  $z_m$  then finally yields Eq. (26).

Note that, the integral over  $z_m$  only converges if

$$\sigma_m^2 \leq \frac{1}{2\beta |\Delta f|} \frac{f_{\lambda(t_f)}}{f_{\lambda(0)}}. \quad (\text{A.11})$$

### Appendix A.3. Derivation of measured Jarzynski equation for a TLS with continuous driving

In this section we derive the analytic expression of the MJE for a driven TLS, namely Eq. (28). We assume that the protocol  $\lambda(t)$  changes continuously and is piecewise differentiable as in Appendix A.1. For the TLS it also holds that  $z^* = z$  and  $y^* = y$ . The work along a trajectory  $\mathbf{z}$  can be discretized as

$$W[\mathbf{z}] = \sum_i (\varepsilon_{\lambda_i}(z_{i-1}) - \varepsilon_{\lambda_{i-1}}(z_{i-1})) \equiv \sum_i \delta \varepsilon_{\lambda_i}(z_{i-1}). \quad (\text{A.12})$$

Equivalently, the measured work is given by  $W_m[\mathbf{y}] = \sum_i \delta \varepsilon_{\lambda_i}(y_{i-1})$ .

Then we can evaluate the right hand side of Eq. (20) analytically as follows:

$$\begin{aligned} & \left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} \\ &= \sum_{\mathbf{z}^\dagger} \mathcal{P}^\dagger[\mathbf{z}^\dagger] \prod_i \left( \sum_{y_i} [(1 - \eta)\delta_{y_i, z_i} + \eta(1 - \delta_{y_i, z_i})] e^{\beta(\delta \varepsilon_{\lambda_{i+1}^\dagger}(y_i) - \delta \varepsilon_{\lambda_{i+1}^\dagger}(z_i))} \right) \\ &= \sum_{\mathbf{z}^\dagger} \mathcal{P}^\dagger[\mathbf{z}^\dagger] \prod_i \left( (1 - 2\eta) + \eta \left[ e^{\beta(\delta \varepsilon_{\lambda_{i+1}^\dagger}(g) - \delta \varepsilon_{\lambda_{i+1}^\dagger}(z_i))} + e^{\beta(\delta \varepsilon_{\lambda_{i+1}^\dagger}(e) - \delta \varepsilon_{\lambda_{i+1}^\dagger}(z_i))} \right] \right). \end{aligned} \quad (\text{A.13})$$

Here,  $\sum_{\mathbf{z}^\dagger} = \sum_{z_N} \dots \sum_{z_0}$  denotes all the sums over  $z_k$  and  $\delta \varepsilon_{\lambda_k^\dagger}(z_{k-1})$  is defined as in Eq. (A.12) with the time-reversed protocol  $\lambda^\dagger(t)$ . To further simplify Eq. (A.13) we introduce the complementary state  $\bar{z}_k$  such that  $\bar{z}_k \neq z_k$  for all  $k$ , i.e. if  $z_k = e$  then  $\bar{z}_k = g$  and vice versa. Consequently,

$$\left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} = \sum_{\mathbf{z}^\dagger} \mathcal{P}^\dagger[\mathbf{z}^\dagger] \prod_i \left( (1 - 2\eta) + \eta \left[ 1 + e^{\beta(\delta \varepsilon_{\lambda_{i+1}^\dagger}(\bar{z}_i) - \delta \varepsilon_{\lambda_{i+1}^\dagger}(z_i))} \right] \right). \quad (\text{A.14})$$

For large  $N$  we approximate

$$1 + e^{\beta(\delta\varepsilon_{\lambda_{i+1}^\dagger}(\bar{z}_i) - \delta\varepsilon_{\lambda_{i+1}^\dagger}(z_i))} \approx 2 + \beta(\delta\varepsilon_{\lambda_{i+1}^\dagger}(\bar{z}_i) - \delta\varepsilon_{\lambda_{i+1}^\dagger}(z_i)) \equiv 2 + \beta\delta_i, \quad (\text{A.15})$$

such that we can write Eq. (A.14) simply as

$$\left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} = \sum_{\mathbf{z}^\dagger} \mathcal{P}^\dagger[\mathbf{z}^\dagger] \prod_i (1 + \eta\beta\delta_i). \quad (\text{A.16})$$

Writing the product explicitly yields

$$\left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} = \sum_{\mathbf{z}^\dagger} \mathcal{P}^\dagger[\mathbf{z}^\dagger] \sum_{n=0}^N \left( \frac{1}{n!} (\eta\beta)^n \sum_{i_1 \neq \dots \neq i_n} \delta_{i_1} \times \dots \times \delta_{i_n} \right). \quad (\text{A.17})$$

We now make the crucial assumption that  $\mathcal{P}^\dagger[z_{k_1}, \dots, z_{k_n}] \approx p^\dagger(z_{k_1}) \dots p^\dagger(z_{k_n})$ . Then,

$$\left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} = \sum_{\mathbf{z}^\dagger} \sum_{n=0}^N \frac{(\eta\beta)^n}{n!} \sum_{i_1, \dots, i_n} p^\dagger(z_{i_1}) \delta_{i_1} \dots p^\dagger(z_{i_n}) \delta_{i_n} - \mathcal{R}. \quad (\text{A.18})$$

To ensure this equality, we introduced a “rest” term  $\mathcal{R}$  of the form

$$\begin{aligned} \mathcal{R} = & \frac{(\eta\beta)^2}{2!} \sum_i \sum_{z_i} p^\dagger(z_i)^2 \delta_i^2 \\ & + \frac{(\eta\beta)^3}{3!} \sum_{ij} \sum_{z_i, z_j} p^\dagger(z_i)^2 \delta_i^2 p^\dagger(z_j) \delta_j + \frac{(\eta\beta)^3}{3!} \sum_i \sum_{z_i} p^\dagger(z_i)^3 \delta_i^3 + \dots \end{aligned} \quad (\text{A.19})$$

taking care of the sums where at least two of the indices  $i_1, \dots, i_n$  are equal. But then all terms of  $\mathcal{R}$  are at least of the order  $\mathcal{O}(\frac{1}{N})$  and therefore vanish for  $N \rightarrow \infty$ . Hence, we are left with evaluating

$$\left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} = \sum_{i=0}^N \frac{1}{n!} (\eta\beta)^n \sum_{i_1, \dots, i_n} \sum_{z_{i_1}, \dots, z_{i_n}} p^\dagger(z_{i_1}) \delta_{i_1} \dots p^\dagger(z_{i_n}) \delta_{i_n}. \quad (\text{A.20})$$

Taking the limit  $N \rightarrow \infty$ , we can write

$$\lim_{\delta t \rightarrow 0} \frac{\delta_k}{\delta t} = \lim_{\delta t \rightarrow 0} \frac{\delta\varepsilon_{\lambda_{k+1}^\dagger}(\bar{z}_k) - \delta\varepsilon_{\lambda_{k+1}^\dagger}(z_k)}{\delta t} = \dot{\varepsilon}_{\lambda^\dagger(t_{k+1})}(\bar{z}_k) - \dot{\varepsilon}_{\lambda^\dagger(t_{k+1})}(z_k) \quad (\text{A.21})$$

where we again assumed that the protocol is differentiable (see the remark below for the case of a c.p.d. protocol). Evaluating the sums over  $z_{i_k}$  and writing the sums over  $i_k$  as integrals (by taking  $N \rightarrow \infty$ ), Eq. (A.20) finally reads

$$\left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} \approx \sum_{n=0}^{\infty} \frac{1}{n!} (-1)^n \left( \eta\beta \int_0^{t_f} dt \dot{\omega}_{\lambda^\dagger(t)} (p_e^\dagger(t) - p_g^\dagger(t)) \right) \quad (\text{A.22})$$

where we denote the time derivative of the energy gap of the TLS by  $\dot{\omega}_{\lambda^\dagger(t_k)} = \dot{\varepsilon}_{\lambda^\dagger(t_k)}(e) - \dot{\varepsilon}_{\lambda^\dagger(t_k)}(g)$  and the probability of the system to be in the ground/exited state

at time  $t_i$  by  $p_{g/e}^\dagger(t_i)$ , both in the backward protocol of the driving scheme. Note that Eq. (A.22) is exact up to first order in  $\eta$ .

Finally, we remark that for a c.p.d. protocol with nondifferentiable points at  $0 < t_1 < \dots < t_K < t_f$  the result above readily generalizes and in Eq. (A.22) we have to split the integral at the respective points as

$$\int_0^{t_f} dt = \int_0^{t_1} dt + \int_{t_1}^{t_2} dt + \dots + \int_{t_K}^{t_f} dt. \quad (\text{A.23})$$

#### Appendix A.4. Derivation of MJE for a TLS for instantaneous driving

In this section we derive Eq. (29), i.e. an expression for the MJE of a TLS, where the energy levels are changed instantaneously at one moment in time  $t_m$  with  $0 < t_m < t_f$  and are constant before and after. Since the energy levels are constant before and after  $t_m$ , it follows that  $\delta\varepsilon_{\lambda_{i+1}^\dagger}(z_i) = 0$  for all  $i \neq m$  and also  $\delta\varepsilon_{\lambda_{i+1}^\dagger}(\bar{z}_i) = 0$  for all  $i \neq m$ . Then the right hand side of Eq. (20) simplifies to

$$\begin{aligned} \left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} &= \sum_{\mathbf{z}^\dagger} \mathcal{P}^\dagger[\mathbf{z}^\dagger] \left[ 1 - \eta \left( 1 - e^{\beta(\delta\varepsilon_{\lambda_{m+1}^\dagger}(z_m) - \delta\varepsilon_{\lambda_{m+1}^\dagger}(\bar{z}_m))} \right) \right] \\ &= \sum_{z_m \in \{g, e\}} p_{z_m}^\dagger(t_m) \left[ 1 - \eta \left( 1 - e^{\beta(\delta\varepsilon_{\lambda_{m+1}^\dagger}(z_m) - \delta\varepsilon_{\lambda_{m+1}^\dagger}(\bar{z}_m))} \right) \right]. \end{aligned} \quad (\text{A.24})$$

Summing over  $z_m$ , Eq. (A.24) can be written as

$$\left\langle e^{\beta(W_m^\dagger - W^\dagger)} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} = 1 - \eta \left( 1 - p_g^\dagger(t_m) e^{\beta\Delta\omega^\dagger} - p_e^\dagger(t_m) e^{-\beta\Delta\omega^\dagger} \right) \quad (\text{A.25})$$

where  $\Delta\omega^\dagger = \omega_{\lambda^\dagger(t_f)} - \omega_{\lambda^\dagger(0)}$ . Note that this equation is exact for  $N \rightarrow \infty$  ( $\delta t \rightarrow 0$ ).

#### Appendix A.5. Derivation of the Brownian particle under feedback

For the derivation of Eq. (39), the MJE under feedback, we assume that  $\mu_{\lambda(0)} = 0$  initially and changes instantaneously at  $t_m$  to  $\mu_{\lambda(t_f)}$  if  $y_m > 0$ . Similarly, the width  $f_{\lambda(t)}$  changes from  $f_{\lambda(0)}$  to  $f_{\lambda(t_f)}$  instantaneously if  $y_m > 0$ . Since the form and the position of the potential is fixed before and after applying the feedback, it holds  $\delta H_{\lambda_{k+1}(y_m)}(y_k) = 0$  for all  $k \neq m$  and the same is true for  $z_k$ . Then the measured efficacy parameter reads after integration over all  $z_k$  and  $y_k$  with  $k \neq m$ :

$$\gamma_m = \int dz_m \int dy_m p_{\lambda^\dagger(y_m)}(z_m) p_m(y_m|z_m) e^{\beta(\delta H_{\lambda_{m+1}^\dagger(y_m)}(y_m) - \delta H_{\lambda_{m+1}^\dagger(y_m)}(z_m))}. \quad (\text{A.26})$$

The integral of  $y_m$  splits into two parts: one in which we alter the potential ( $y_m > 0$ ) and one where we do nothing ( $y_m < 0$ ):

$$\begin{aligned} \left\langle e^{\beta(W_m^\dagger[\mathbf{y}^\dagger|y_m] - W^\dagger[\mathbf{z}^\dagger|y_m])} \right\rangle_{\mathbf{z}^\dagger, \mathbf{y}^\dagger} &= \int dz_m \int_{-\infty}^0 dy_m p_{\lambda^\dagger(y_m)}(z_m) p_m(y_m|z_m) \\ &+ \int dz_m \int_0^\infty dy_m p_{\lambda^\dagger(y_m)}(z_m) p_m(y_m|z_m) e^{\beta(\delta H_{\lambda_{m+1}^\dagger(y_m)}(y_m) - \delta H_{\lambda_{m+1}^\dagger(y_m)}(z_m))}. \end{aligned} \quad (\text{A.27})$$

The conditional probability  $p_m(y_m|z_m)$  is again assumed to be Gaussian with a standard deviation of  $\sigma_m$  (see Eq. (24)). Moreover, the probability  $p_{\lambda^\dagger(y_m < 0)}(z_m)$  (no feedback) is the canonical distribution of the harmonic potential centered at  $\mu_{\lambda(0)}$  and width  $f_{\lambda(0)}$  and the probability  $p_{\lambda^\dagger(y_m \geq 0)}(z_m)$  (feedback) is the canonical distribution centered at  $\mu_{\lambda(t_f)}$  and width  $f_{\lambda(t_f)}$ , because we are in equilibrium before applying the backwards protocol. Then the first term of Eq. (A.27) becomes 1/2 after integration of  $z_m$  and  $y_m$ . If feedback is applied ( $y_m > 0$ ) it holds

$$\delta H_{\lambda_{m+1}^\dagger(y_m)}(y_m) - \delta H_{\lambda_{m+1}^\dagger}(z_m) = (f_{\lambda(0)} - f_{\lambda(t_f)})(y_m^2 - z_m^2) - 2f_{\lambda(t_f)}\mu_{\lambda(t_f)}(z_m - y_m). \quad (\text{A.28})$$

Then after integration over  $z_m$  and  $y_m$  of the second part of Eq. (A.27) one arrives at Eq. (39).

#### Appendix A.6. Derivation for the two level system under feedback

Here, we derive the analytic expression of the MJE for a driven TLS under feedback, Eq. (45). We again assume that the driving protocol changes continuously and depends on the measurement outcome  $y_m$  at time  $t_m$ . Then, the measured efficacy parameter of the TLS is given by

$$\gamma_m = \sum_{\mathbf{z}^\dagger} \sum_{\mathbf{y}^\dagger} \mathcal{P}_{\lambda^\dagger(y_m)}[\mathbf{z}^\dagger] \prod_k [(1 - \eta)\delta_{y_k, z_k} + \eta(1 - \delta_{y_k, z_k})] e^{\beta(\delta\varepsilon_{\lambda_{k+1}}(y_m)(y_k) - \delta\varepsilon_{\lambda_{k+1}}(y_m)(z_k))}. \quad (\text{A.29})$$

Since the driving protocol depends on  $y_m$ , we can write  $\gamma_m$  as:

$$\begin{aligned} \gamma_m &= \sum_{\mathbf{z}^\dagger} \sum_{y_m} \mathcal{P}_{\lambda^\dagger(y_m)}[\mathbf{z}^\dagger] [(1 - \eta)\delta_{y_m, z_m} + \eta(1 - \delta_{y_m, z_m})] e^{\beta(\delta\varepsilon_{\lambda_{m+1}}(y_m)(z_m) - \delta\varepsilon_{\lambda_{m+1}}(y_m)(y_m))} \\ &\times \prod_{k \neq m} \left[ \sum_{y_k} (1 - \eta)\delta_{y_k, z_k} + \eta(1 - \delta_{y_k, z_k}) e^{\beta(\delta\varepsilon_{\lambda_{k+1}}(y_m)(z_k) - \delta\varepsilon_{\lambda_{k+1}}(y_m)(y_k))} \right]. \end{aligned} \quad (\text{A.30})$$



Summing over all  $y_k$  results in

$$\begin{aligned}
\gamma_m &= \sum_{\mathbf{z}^\dagger} \sum_{y_m} \mathcal{P}_{\lambda^\dagger(y_m)}[\mathbf{z}^\dagger] [(1-\eta)\delta_{y_m, z_m} + \eta(1-\delta_{y_m, z_m})] e^{\beta(\delta\varepsilon_{\lambda_{m+1}}(y_m)(z_m) - \delta\varepsilon_{\lambda_{m+1}}(y_m)(y_m))} \\
&\quad \times \prod_{k \neq m} \left( (1-2\eta) + \eta \left[ 1 + e^{\beta(\delta\varepsilon_{\lambda_{i+1}^\dagger}(\bar{z}_i) - \delta\varepsilon_{\lambda_{i+1}^\dagger}(z_i))} \right] \right) \\
&\approx \sum_{\mathbf{z}^\dagger} \sum_{y_m} \mathcal{P}_{\lambda^\dagger(y_m)}[\mathbf{z}^\dagger] [(1-\eta)\delta_{y_m, z_m} + \eta(1-\delta_{y_m, z_m})] \\
&\quad \times \prod_{k \neq m} \left( (1-2\eta) + \eta \left[ 1 + e^{\beta(\delta\varepsilon_{\lambda_{i+1}^\dagger}(\bar{z}_i) - \delta\varepsilon_{\lambda_{i+1}^\dagger}(z_i))} \right] \right).
\end{aligned} \tag{A.31}$$

For the last step we approximated

$$\exp \left\{ \beta(\delta\varepsilon_{\lambda_{m+1}}(y_m)(y_m) - \delta\varepsilon_{\lambda_{m+1}}(y_m)(z_m)) \right\} \approx 1 \tag{A.32}$$

for the single point at  $k = m$ . This is justified because the final integral does not depend on the value of a single point as long as we change the protocol continuously. Following the same intermediate steps as in Appendix A.3 we arrive at

$$\begin{aligned}
\gamma_m &\approx \sum_{z_m, y_m} [(1-\eta)\delta_{y_m, z_m} + \eta(1-\delta_{y_m, z_m})] [p_{\lambda^\dagger(y_m)}(z_m) \\
&\quad \times \exp \left\{ -\beta\eta \int dt \dot{\omega}_{\lambda(y_m, t)}^\dagger (p_{e, \lambda^\dagger(y_m)}(t) - p_{g, \lambda^\dagger(y_m)}(t)) \right\}] .
\end{aligned} \tag{A.33}$$

Finally, by summing over  $y_m$  we arrive at Eq. (45).

For an instantaneous change of the driving protocol, where we assume that the Hamiltonian of the TLS is constant before and after the quench at time  $t_m$ ,  $\delta\varepsilon_{\lambda_{i+1}}(y_m)(z_m)$  and  $\delta\varepsilon_{\lambda_{i+1}}(y_m)(y_m)$  are the only terms different from zero. Then, Eq. (46) follows immediately from evaluating the sum over  $y_m$  in Eq. (A.29).